DOCUMENT RESUME

ED 178 871 CS 005 107

AUTHOR	Schreiner, Robert, Ed.
TITLE	Reading Tests and Teachers: A Practical Guide.
INSTITUTION	International Reading Association, Newark, Del.
PUB DATE	79
NOTE	90p.; Figure 4.4 may not reproduce
AVAILABLE FROM	International Reading Association, 800 Barksdale Road, P.O. Box 8139, Newark, DE 19711 (Order No. 726, $2.50 member, $4.00 non-member)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Evaluation Methods: Reading Ability; *Reading Achievement; *Reading Diagnosis; *Reading Tests; *Test Construction; *Testing; Test Interpretation; *Test Selection

ABSTRACT

This guidebook puts testing in proper perspective as one aspect of evaluation and leads the test user through a practical discussion about selecting and developing valid, reliable tests. Its primary purpose is to assist classroom teachers, reading consultants, and school administrators in making decisions about measuring student reading achievement. The five chapters of the book discuss the following topics: a plan for using tests in reading evaluation, how to select and evaluate reading tests, the construction and analysis of reading tests, how to organize and apply test results, and definitions of measurement terms. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178871

Reading Tests and Teachers: a practical guide

Robert Schreiner, Editor
University of Minnesota

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

International Reading Association

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS005107

INTERNATIONAL READING ASSOCIATION
800 Barksdale Road • Newark, Delaware 19711

INTERNATIONAL READING ASSOCIATION

OFFICERS
1979-1980

President Roger Farr, Indiana University, Bloomington, Indiana

Vice President Olive S. Niles, State Department of Education, Hartford, Connecticut

Vice President-Elect Kenneth S. Goodman, University of Arizona, Tucson, Arizona

Executive Director Ralph C. Staiger, International Reading Association, Newark, Delaware

DIRECTORS

Term Expiring Spring 1980
Mary Ann Baird, State Department of Education, Jackson, Mississippi
Addie Stabler Mitchell, Morehouse College, Atlanta, Georgia
Jean E. Robertson, University of Alberta, Edmonton, Alberta

Term Expiring Spring 1981
Norma Dick, Clovis Unified School District, California
Eleanor M. Ladd, University of South Carolina, Spartanburg, South Carolina
John C. Manning, University of Minnesota, Minneapolis, Minnesota

Term Expiring Spring 1982
Nicholas P. Criscuolo, Betsy Ross Middle School, New Haven, Connecticut
Bernice E. Cullinan, New York University, New York, New York
Patricia S. Koppman, San Diego Unified Schools, San Diego, California

Copyright 1979 by the
International Reading Association, Inc.

Library of Congress Cataloging in Publication Data
Main entry under title:

Reading tests and teachers.

Bibliography: p.
1. Reading—Ability testing. I. Schreiner, Robert, 1937- II. International Reading Association. Evaluation of Tests Committee.
LB1050.46.R4 428'.4'076 79-17271
ISBN 0-87207-726-8

Contents

Foreword *v*

Acknowledgements *vi*

Evaluation of Tests Committee *vii*

Table of Figures *viii*

1 Chapter 1 TESTING AND READING: A PLAN FOR EVALUATION Robert Schreiner
- Domain of Reading Behaviors *2*
- Goals for Reading Instruction *2*
- Assessment of Needs *3*
- Specific Instructional Objectives *4*
- Instructional Program *6*
- Measurement and Evaluation of Performance *6*

9 Chapter 2 SELECTING AND EVALUATING READING TESTS Kenneth Kavale
- Factors in Test Selection *11*
  - Validity *11*
  - Reliability *13*
  - Usability *13*
  - Norm Information *14*
- Test Selection: The Final Process *15*
  - Test Selection Checklist *15*
- What Types of Tests are Available? *17*
  - Standardized Norm Referenced Reading Achievement and/or Survey Tests *17*
  - Criterion Referenced Testing *21*
  - Informal Reading Measures *24*
  - Diagnostic Reading Tests *28*
  - Reading Readiness Tests *30*

35 Chapter 3 CONSTRUCTION AND ANALYSIS OF READING TESTS Michael L. Kamil
Reading Readiness Skills 36
Decoding Skills 38
Structural Analysis Skills 38
Vocabulary Skills 42
Comprehension Skills 43
Initial Considerations 45
Constructing Objectives and Test Items 46
Writing Objectives 47
Writing Items 49
The Finishing Touches 51

53 Chapter 4 ORGANIZING AND APPLYING TEST RESULTS George Canney
How to Use Test Scores 53
Validate 54
Equilibrium 55
Review 58
Investigate 63
Formulate 68
Interpret 68
Using Assessment to Guide Instruction 69

72 Chapter 5 DEFINITIONS OF MEASUREMENT TERMS Fred Pyrczak
Types of Tests 73
Norm Referenced Test 73
Criterion Referenced Test 73
Domain Referenced Test 74
Diagnostic Test 74
Survey Test 75
Group Test 75
Individual Test 75
Types of Scores 75
Percentile Rank 75
Percentile Band 76
Grade Equivalent 76
Stanine 77
Essential Characteristics of Tests 77
Validity 77
Reliability 80
Standard Error of Measurement 80

Foreword

Gathering information regarding students' reading achievements is an integral part of instruction, and one means of gathering such information is through the use of tests. Testing should, therefore, be a normal and well understood activity; but that is often not the case. Because of the technical vocabulary used in testing, the aura of sophistication of tests, and an increasing reliance on tests (particularly in state minimum competency programs), testing has become a seemingly complex activity.

The International Reading Association's Committee on the Evaluation of Tests has attempted to reduce the complexity and confusion surrounding tests and their uses. In previous years, the committee has published user oriented reviews of tests. Edited by Robert Schreiner, this monograph on *Reading Tests and Teachers: A Practical Guide* expands on previous committee publications and should take some of the mystery out of testing.

The Committee has developed this publication as a guide which first puts testing in proper perspective as one aspect of evaluation. It then leads the test user through a practical discussion of how to select tests and how to develop valid and reliable tests. Once a test has been selected and administered, the results should be used to plan instruction; and the guide includes a very functional approach to using test scores.

Competency testing, criterion referenced measurement, accountability, and a host of other test and test related concepts face reading teachers every day. This practical guide will help test users through the morass of jargon and help them to make tests a tool for planning instruction rather than a device that dictates the curriculum.

Roger Farr, *President*
International Reading Association
1979-1980

Acknowledgements

This guide to the construction and use of tests and the interpretation of test results is a project of the International Reading Association Evaluation of Tests Committee.

The committee is deeply indebted to many individuals for this final product. Initially an outline of the contents was prepared following an extensive survey of assessment and evaluation needs of teachers, school administrators, reading consultants, and college and university professors. Authors were contacted and writing tasks agreed upon. Extensive authorial review and rewriting occurred before the first draft was critically reviewed. Rewriting followed the initial review and the final manuscript was edited.

I am particularly indebted to George Canney, Michael Kamil, Kenneth Kavale, and Fred Pyrczak for their perseverance and dedication to this project. Leo M. Schell (Kansas State University), provided extensive editing and review expertise. Clarence Brock (St. Clair County, Michigan, Public Schools), Ronald Carver (University of Missouri at Kansas City), and Roger Farr (Indiana University), aided us with extensive constructive criticism. Finally Walter MacGinitie (Teachers College, Columbia University), Ralph C. Staiger (Executive Director, International Reading Association), and the members of the Evaluation of Tests Committee provided constant professional support for this effort.

Without all these dedicated professionals, this project would not have materialized.

Robert Schreiner, *Editor*

IRA PUBLICATIONS COMMITTEE 1979-1980 Harold L. Herber, Syracuse University, *Chairing* • Alison Bellack, Largo (Florida) C & I Center • Janet R. Binkley, IRA • Faye R. Branca, IRA • Margaret Keyser Hill, Southern Illinois University • Jerry L. Johns, Northern Illinois University • Lloyd W. Kline, IRA • J. E. Merritt, The Open University • Addie S. Mitchell, Morehouse College • Christina Neal, St. Albans, West Virginia • P. David Pearson, University of Illinois at Champaign • Emma W. Rembert, Florida International University • James D. Riley, University of Montana • Maria Elena Rodriguez, Asociacion Internacional de Lectura, Buenos Aires • Robert B. Ruddell, University of California at Berkeley • Carol Sager, Chestnut Hill, Massachusetts • S. Jay Samuels, University of Minnesota • Lyndon W. Searfoss, Arizona State University • Cyrus F. Smith, Jr., University of Wisconsin at Milwaukee • Ralph C. Staiger, IRA • Judith Thelen, Frostburg (Maryland) State College • Carl J. Wallen, Arizona State University.

The International Reading Association attempts, through its publications, to provide a forum for a wide spectrum of opinion on reading. This policy permits divergent viewpoints without assuming the endorsement of the Association.

EVALUATION OF TESTS COMMITTEE
1977-1978

Robert Schreiner, *Chairing*
University of Minnesota at Minneapolis

Jean E. Robertson, *Board Liaison*
University of Alberta

George Canney
University of Idaho

Ronald P. Carver
University of Missouri at Kansas City

Warwick Elley
University of South Pacific

Roger Farr
Indiana University

Sherry Gable
University of Northern Iowa

Vincent Greaney
St. Patrick's College
Dublin, Ireland

Martha Rupp Haggard
Northern Illinois University at DeKalb

Karen M. Hanson
Eden Prairie, Minnesota

Michael L. Kamil
Purdue University

Kenneth Kavale
University of California at Riverside

Fred Pyrczak
California State University at Los Angeles

Leo M. Schell
Kansas State University

TABLE OF FIGURES

Figure 1.1	Steps for Making Diagnostic and Instructional Decisions in Assessing Reading Achievement	*2*
Figure 3.1	Sample, Cooper-McGuire Diagnostic Word Analysis Test item	*37*
Figure 3.2	Sample, Cooper-McGuire Diagnostic Word Analysis Test item	*37*
Figure 3.3	Sample, Stanford Diagnostic Reading Test item	*39*
Figure 3.4	Sample, Stanford Diagnostic Reading Test item	*40*
Figure 3.5	Sample, Stanford Diagnostic Reading Test item	*40*
Figure 3.6	Sample, Cooper-McGuire Diagnostic Word Analysis Test item	*41*
Figure 3.7	Sample, Gates-McKillop Reading Diagnostic Tests item	*42*
Figure 3.8	Sample, Slosson Oral Reading Test item	*42*
Figure 3.9	Sample, Johnson Basic Sight Vocabulary Test item	*43*
Figure 3.10	Sample, Davis Reading Test item	*43*
Figure 4.1	An Item Response Chart Showing Correct and Incorrect Responses on a Test of Prefixes	*55*
Figure 4.2	Code System for Oral Reading for Diagnosis	*59*
Figure 4.3	Every Pupil Response Devices for Maximizing Practice in Reading	*61*
Figure 4.4 a, b, c, d	Pupil Progress Chart	*64*

Chapter One

Testing and Reading: A Plan for Evaluation

Robert Schreiner
University of Minnesota at Minneapolis

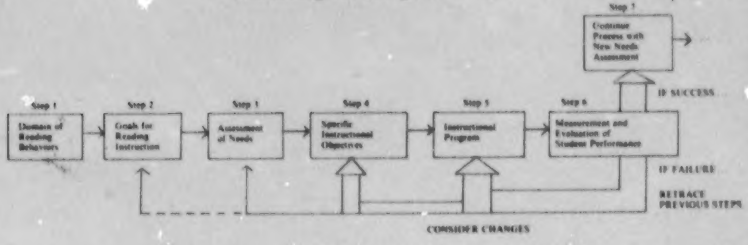
The primary purpose of this publication is to assist classroom teachers, reading consultants, and/or school administrators in making decisions about measuring the reading achievement of students.

The use of tests in any total reading program is intimately linked with decision making. School personnel are constantly faced with a variety of decisions about reading instruction: What are the appropriate goals of instruction? What kinds of instructional needs do my students have? What skills should be taught? How might the skills be translated into instructional objectives? Which materials will best aid instruction? What grouping plan is best? Who is making progress? Is more help needed? The questions may be never ending, but the answers must be arrived at rationally.

A variety of published tests is available to assist in making instructional decisions. In addition, many informal testing techniques may be used.

The following flow chart suggests a set of systematic procedures to use in making decisions regarding student performance in reading.

Figure 1.1
Steps for Making Diagnostic and Instructional Decisions
in Assessing Reading Achievement



Step 1 Domain of Reading Behaviors →

Examining the Domain of Reading Behaviors

The domain of reading behavior includes all the perceptual, linguistic and cognitive skills we bring to printed words to pronounce them and understand them, including our recognition and understanding of letters, letter sound relationships, decoding of whole words, phrases and sentences, the grammatical aspects of our language, and how we obtain meaning from print. All the aspects of written language, of course, are not related to the reading act. We are only interested in those behaviors people use and need in the reading act. Behavioral theorists have identified these tasks for us. We only need to think a moment to recall those aspects of visual and auditory perception, cognition, and learning from psychology textbooks that are relevant for reading. These components also include affective behaviors, such as motivation and interest that relate to the act of reading. The domain of reading behaviors can also be referred to as the constructs of the reading act.

→ Step 2 Goals for Reading Instruction →

Determining the Goals for Reading Instruction

The goals for teaching students how to read reflect how we translate the domain of reading behaviors into instructional

practices. This translation normally takes place in school. We teach students *how* to read. That is, reading instruction is both systematic and developmental. For example, at the appropriate time, we provide phonics or decoding lessons based on what we know about symbol/sound correspondence. We instruct students to read whole words, phrases, and sentences with understanding. Comprehension and vocabulary building activities are based on how we think young people learn and use syntactic and semantic information. Instructional strategies are deliberate and planned. These plans are not haphazard and they are based on our knowledge of the constructs involved in reading.

Both the domain of reading behaviors and the development of goals for reading instruction are crucial aspects of assessment and evaluation. If there are serious gaps in our knowledge of cognitive processes or in our translation of that information to sound reading practices, the outcome of assessment may be questioned. Thus, it is important to learn as much as possible about language and thinking and how it relates to the teaching of reading. In other words, our practices must be based on sound theory.

→ *Step 3 Assessment of Needs* →

Assessment begins when we measure or test student performance relative to the established goals for reading instruction. Assessment of needs seeks answers to two complementary questions: How well do students presently read? and What do they need to know to improve their reading performance?

In seeking answers to the above questions, we may wish to use existing background information on students, including information found in the permanent school records. While students should not be classified solely by previous test scores, such information can be used as an indication of instructional needs. The major points of information for us include 1) the reading achievement level estimated from the achievement test records for the previous year and 2) major strengths and weak-

nesses in reading ability as recorded by previous teachers. For these reasons, testing is a crucial part of a needs assessment. Through tests, we obtain information useful for planning instructional strategies. We gain information regarding a student's progress, placement, and readiness or we may diagnose reading strengths and/or weaknesses.

Through a variety of tests, we can accumulate information about the abilities of a class. We are now in a position to make statements about student performance and needs. We should focus upon answers to the following questions:

1. What is the student's current level of performance in word recognition, vocabulary meaning, and reading comprehension?
2. What additional instruction would aid students to perform at a higher level?
3. What combination of instructional program, materials, and techniques is most suitable for student needs?

When we compile this information for an individual or group of students, assessment of needs is well underway.

→ *Step 4* *Specific Instructional Objectives* →

Once we know "where they are" we deal with the question of "where are they going?" We can develop instructional goals based on the results of a needs assessment. Barrett (*1*) suggested three categories of interrelated and interdependent goals for reading programs:

Categories of Goals of the Reading Program

Perceptual Goals	Cognitive Goals	Affective Goals
are concerned with the flexibility rate and accuracy of word perception	are concerned with the types and levels of thought generated during reading	are concerned with the feelings children have toward reading

Depending upon the existing needs, each goal will be emphasized to different degrees.

The goals for the reading program should lead to specific instructional objectives. Objectives give direction to the educa-

tional process by aiding you in 1) making students aware of the goals to be achieved; 2) planning learning activities designed to achieve various goals; 3) organizing time, space, and materials for carrying out learning activities; and 4) providing a criterion for evaluating performance.

Consider the following three factors regarding goals and objectives when selecting a reading test.

First, the final level to be attained depends on the initial status of the student or students. The goals and objectives may be expressed in terms of growth (a year's growth in reading for a year's instruction) or they may be expressed in terms of expectancy (expected reading performance for various ability levels).

Second, how directly or indirectly do the goals and objectives specify reading skills? A direct goal refers to specific reading skills (locating supporting details, identifying "long" vowel sounds, or differentiating singular from plural forms of words). An indirect goal refers to behavior which is related to reading skills but does not directly involve the skill (enjoyment of reading or reading for pleasure).

Third, how are the goals and objectives differentiated in terms of the content of the curriculum? The following categories of major goals and objectives of reading instruction appear in most college methods of teaching reading textbooks:

Decoding skills (word attack or phonics); vocabulary (pronunciation and meaning); structural analysis skills (syllabication, root words, inflected endings, prefixes, and suffixes); oral reading skills (pronunciation, phrasing, and appropriate voice inflection); silent reading; reading rate; reading comprehension (literal and inferential); reading study skills; critical and evaluative reading abilities; affective components such as motivation and student interest; reading in the content areas; and ability to recognize literary devices such as mood, plot, characterization, and setting.

Decisions regarding the point and degree of emphasis of these objectives need to be made in relation to the reading materials used and the age and abilities of the readers.

→ *Step 5 Instructional Program* →

This component of instructional decision making refers to the actual way in which teaching is done. It includes the instructional materials, the delivery of the content to the student, and also the setting in which the instruction is provided.

We need to determine which method of instruction is most appropriate. We may need to use various methods until we find the most efficient or useful procedure. For example, if we find that a student has serious deficits in decoding skills, we may wish to begin instruction with a single-letter phonics approach. Thus, our instruction would be directed toward learning symbol/sound relationships and blending these sounds into words. This method would also suggest that instruction take place in a one-to-one or small group setting. Similar plans would be appropriate if we prescribed a whole word or sight word method for teaching word attack skills. In this case, students would learn whole words and then later analyze new words, using previous knowledge or skills. Instruction in whole word recognition skills is frequently provided in a group setting. We may further consider a combination method that utilizes components of both a structured phonics and whole word method for teaching word attack skills.

The instructional program, as you can see, is crucial in the overall sequence for making instructional decisions. We need to learn as much as possible about the various methods and techniques for all the goals of reading listed earlier as well as how these can best be taught.

→ *Step 6 Measurement and Evaluation of Performance* →

With the goals and objectives of the reading program established, we then focus upon actual instruction. After a period of time, the final step in the assessment sequence must be considered: Evaluation of student performance with respect to progress towards achieving the stated goals and objectives of the reading program. The essential concern is to identify discrepan-

cies, if any, between "what is" and "what should be." If the desired goals have been attained, we can then formulate new objectives for the students. If the desired goals have *not* been attained, it is suggested that we reexamine the instructional component. We may wish to continue with the same instructional program or we may wish to make changes in it.

Sometimes we may want to reexamine the specific instructional objectives with the intent of making revisions. This should happen less often than the revision of instructional programs. Still less frequent should be the reexamination of needs assessment. A faulty needs assessment, in some instances, can be the cause of irrelevant instruction and consequent failure of students to learn to read.

In summary, the steps for making diagnostic and instructional decisions in assessing reading achievement (Figure 1.1) should be followed systematically to complete the evaluation process. The evaluation process is completed with Step 6 (Measurement and Evaluation of Student Performance). If success is achieved, new instructional objectives are presented and the cycle continues. If failure is encountered, we must retrace the steps to determine what caused the failure. As the flow chart suggests (Figure 1.1) we might return to the steps and ask the following questions:

1. Did we use appropriate instruments and/or evaluative techniques to assess student performance? (Step 6)
2. Was the instructional program sensitive enough to affect student progress? (Step 5)
3. Were the instructional objectives appropriate? (Step 4)
4. How satisfactory (reliable and/or valid) were the instruments upon which the original needs assessment were based? (Step 3)
5. Were the goals for reading instruction well delineated? (Step 2)

As can be seen, any breakdown in this chain of events can cause difficulties in the evaluation process.

The remainder of this publication deals more specifically with the steps for making instructional decisions related to

measuring reading progress. In Chapter Two, suggestions are made to aid you in selecting appropriate commercially prepared tests that measure reading progress. Discussion is also directed toward strengths and weaknesses of various kinds of reading tests. Chapter Three provides specific examples of the construction of tests that measure reading progress. Discussion is directed toward the usefulness of both norm and criterion referenced test items. The content of Chapter Four is directed toward classroom uses of the results of measurement and evaluation. Chapter Five presents definitions of the various aspects of the discipline of measurement and evaluation. It also includes a glossary of technical terms used in measurement and evaluation.

Reference

1. Barrett, T.C. "Goals of the Reading Program: The Basis for Evaluation," in T.C. Barrett (Ed.), *The Evaluation of Children's Reading Achievement*. Newark, Delaware: International Reading Association, 1967.

Chapter Two

Selecting and Evaluating Reading Tests

Kenneth Kavale
University of California at Riverside

The major purpose of this chapter is to aid the teacher in answering the question: "Which reading test should I use?" A test should not be used because it is the only one available or because it is the one which has always been used. Old tests sometimes remain in the school long after the curriculum has changed. Such tests may become outdated and no longer serve the original purpose. However, some of the older tests may be among the best available. Consequently, test selection requires time and careful evaluation.

The purposes of testing in reading are quite diverse and no single test can accomplish all these purposes. Different types of tests are implied by each purpose. Therefore, when deciding which test(s) to select, the first consideration should be a detailed examination of the purposes for which the testing is to be done. *If the uses for the test results are not known in advance, the best test to use is none at all.*

Merwin (*15*) has proposed a useful guide for decisions related to test selection. It bases test selection upon a scheme which inquires about the testing of *what*, *who*, by *whom*, to *what*. Below are some examples of how the guide may be used with respect to reading assessment.

1. Assessing reading achievement of a third grade student by a teacher to aid in selecting reading materials.

2. Assessing reading readiness of a group of kindergarten students by a teacher to aid in forming instructional groups.

3. Assessing word recognition skills of a second grade student by a teacher to decide whether a supplementary phonics program is necessary.

4. Assessing gain in reading achievement of a group of fifth grade students by a curriculum supervisor to determine the effectiveness of a new set of instructional materials.

5. Assessing levels of inferential reading comprehension of a tenth grade student by a teacher to determine understanding.

6. Assessing reading performance of a group of eighth grade students by a reading support teacher to decide who needs supplemental instruction.

These examples reveal a portion of the diverse purposes of testing in reading. Therefore, it is essential for the teacher to realize that the reason for measurement will be the primary factor in determining the methods of measurement, that is, which test to choose.

What tests are available? Once the teacher has identified the aims of testing, the next task for the teacher is to locate suitable tests. Perhaps the best single source is *The Eighth Mental Measurements Yearbook* (*5*) or its companion, *Reading: Tests and Reviews* (*4*), with reviews on nearly all published reading tests. Many reading instruction textbooks also list numerous tests and publishers. Basically, the teacher seeks answers to the following questions:

1. Which tests might serve my present purpose?
2. What are the new tests in the field?
3. What is test X like?
4. What do specialists in the field of measurement have to say about test Y?

From among the many tests currently available, several will typically appear to be appropriate for the purpose(s) identified. Choosing from among the identified available tests is the next problem for the teacher.

The decision as to which test is best should be based upon the technical properties of the tests. This is not to suggest that teachers need to be measurement experts but rather that issues such as the content measured by the test, the consistency of the obtained scores, the validity of subtest and total test scores, the adequacy of the normative group (when appropriate), the adequacy of the test manual, and the ease of administration and scoring are important considerations if the best test is to be chosen. (See Chapter Five for more detailed discussion of measurement terms.) A teacher knowledgeable about the strengths and weaknesses of particular tests is far more likely to select the best instrument available for the specified purpose.

Factors in Test Selection

Two major facets of measurement are important for the teacher to consider in test selection: validity and reliability.

Validity

The first and foremost question to be asked with respect to a test instrument is: "Does the test measure what it is intended to measure?" The primary concern for the teacher is whether testing has an effect upon instructional practice and materials. Can the obtained data be used to make significant changes in instruction or by making materials more focused and effective?

Two kinds of validity are important for teachers to consider.

1. *Content validity* which assesses whether the test measures the content being taught. Consider a test prepared to measure reading achievement. How can we tell how well the test does, in fact, measure reading achievement? Three factors are important to consider:

   a. First, there must be some agreement as to the skills, knowledge, and understandings which comprise your definition of reading achievement.
   b. Second, it is necessary to examine the test to determine what skills, knowledge, and understandings are required for it.

c. Finally, it is necessary to determine the match which exists between test content and course content (instructional objectives). To the extent that the course content is represented in the test, the test is considered to possess adequate content validity.

An excellent way for teachers to determine content validity is to take the test. Then they may be in a better position to pass judgment about just what the test is really testing and to determine whether what is being tested matches what is being taught in the classroom.

2. *Face validity*, sometimes considered, refers to the fact that the test looks as if it should be valid. It is important for motivational purposes, for students are more likely to cooperate if they sense that the nature of the items logically correspond to the overall purpose of the test. For example, basal reader tests can be considered to have high face validity. However, the presence of face validity does not guarantee that the test is valid in any other sense.

It would be unwise to assume that the name or title of a test reveals what it measures. An examination of various tests of reading comprehension will reveal that certain tests require the student to determine the main idea of a paragraph; others demand only the retrieval of literal meaning; and a few ask the reader to discern the intent and mood of the author. Tests of reading rate vary in time from 60 seconds to much longer periods of reading. Some give credit for speed, even if many questions are not answered correctly; others give no credit for rate unless comprehension occurs. Some vocabulary tests are constructed of items listing a word with several possible synonyms from which to select the correct answer; in others, the examiner reads a sentence and the student is required to select one of several words to complete the sentence. In still a third type, an underlined word in a sentence is presented and the student responds by marking one of several possible synonyms. It is quite possible that some students will perform better on one type of test for comprehension, rate, or vocabulary than on others; yet, all may have identical titles.

Reliability

While validity ranks as an important attribute of any test, reliability is another important facet to be considered in test selection. Reliability refers to how consistently a test measures whatever it does measure.

Test results should not reflect chance scores with a child's obtaining a high score by luck, guessing, or other extraneous factors. Any factor that tends to exert a varying influence on test scores has an adverse effect on the reliability of a test. Such factors as examinee differences, influence of the scorer, testing situation, time influences, and test content can affect scores differentially, which increases the amount of error and, consequently, lowers the reliability. In examining a test, it is usually the case that reliability will be greater for scores which are:

1. from a longer test than from a shorter test,
2. from a test composed of more homogeneous items than from a more heterogeneous test,
3. from a test whose items are of middle difficulty than from a test composed mainly of quite difficult or quite easy items,
4. obtained from a group having a wide range of ability than from a group more homogeneous in ability.

While validity and reliability are somewhat related, they are by no means identical. Validity refers to the degree to which a test actually serves the purpose for which it is intended; reliability refers to the test's capacity to yield consistent information regardless of whether it serves the purpose or purposes for which it is intended and, in doing so, provide consistent information.

Usability

In addition to the psychometric qualities of validity and reliability, another dimension to be considered in test selection is usability. This refers to a variety of practical considerations such as ease of administration, convenience of scoring, economy, and adequacy of the test manual.

Consider two reading tests, A and B. Test A possesses very high reliability and validity. Test B possesses a lower but adequate level of validity and reliability. There seems to be little question that Test A is the preferred instrument. However, what if Test A required 3 hours to administer, was given individually, took an hour to score, cost 50 cents per student, and could only be administered by a person who has received extensive training? On the other hand, Test B could be administered to a group in one hour, has reusable test booklets, has answer sheets which cost 10 cents each, and one test could be scored in 5 minutes. The decision is now reversed and there is little question that Test B would be preferred.

Thus, usability determines whether a test is practical for widespread use. The question arises as to how the teacher can determine the usability of a test. There is no better way to learn about problems in administration, scoring, appropriateness of content, and adequacy of the test manual than to take the test and then give it to a few children. It should be realized, however, that there is no point in testing unless there is some confidence in the results. While practical factors should be considered, they are important only if the test has satisfactory reliability and validity.

Norm Information

A final important consideration in test selection, when test results are to be used for comparisons of individuals (norm referenced measurement), is the adequacy of the normative information supplied. (A discussion of an alternative to norm referenced measurement, criterion referenced measurement, appears in a later section.)

Norms should be evaluated before they are used. The most important criterion for the teacher to consider before using the norms supplied is: Can the students in my class be legitimately considered equivalent to the population represented by the standardized group? If the answer is yes, then the test would appear to be appropriate, other things being equal. But if the answer is no, then it appears questionable whether the test is appropriate for that particular group of students since the comparisons made may be inappropriate and/or unfair.

Test Selection: The Final Process

Various factors have been presented which should be considered in test selection. The importance of the task cannot be overemphasized because tests must be selected according to specific criteria to prevent relatively superficial factors from exerting too great an influence upon the selection process. A test should never be chosen on the basis of its reputation nor should its contents be judged from the title alone. An examination of the test itself is the only way for the teacher to determine what and how the test is measuring and, consequently, decide whether it will yield the required information.

To aid in compiling the required information, the following test selection checklist may be used as a guide. The teacher may follow the outline for each test to be considered and when completed, the checklists can be used for comparisons among the tests. The teacher is then in a position to choose the best instrument for the desired purposes.

Test Selection Checklist

I. What is your purpose for evaluating this test?
   A. Why do I need a test?
   B. What information do I require?

II. Test Overview
   A. Test information
      1. Name of test—what does name imply about purpose?
      2. Author.
      3. Publisher.
      4. Date of publication and last revision.
   B. Target population
      1. Age level—appropriate for your situation?
      2. Grade level—appropriate for your situation?
      3. Individual or group administration.
      4. Number of available forms—is another form available for future use?
   C. What construct or quality is this test designed to measure?

II. Practical features
   A. Time considerations
      1. Time required to administer total test.
      2. Could the time requirement be handled in your situation?
      3. Time required to score—how involved and time consuming is the scoring procedure?
      4. Who could score the test in your situation? Does scoring involve judgment or simply tallying right and wrong responses?
      5. Is machine scoring available?
   B. Financial considerations
      1. Cost of test booklet.
      2. Cost of answer sheets
      3. Cost of test manuals.
      4. Cost of replacement materials.
      5. Cost of scoring services.
   C. General considerations
      1. Are the directions for administering clear and is the language appropriate for the students in your class?
      2. Are the directions for scoring clear?
      3. Are answer sheets appropriate for age and ability level of students?
      4. What training is required to administer the test?
   D. Format design
      1. Are illustrations current?
      2. Is print size adequate?
IV. Validity
   A. What evidence of validity is given?
   B. Does reported validity appear adequate in relation to intended purposes?
   C. Subtest validity
      1. Does the name adequately describe the subject?
      2. Does the subtest seem logically related to the test as a whole?
V. Reliability
   A. What evidence for reliability is given?
   B. Do the reported reliabilities appear adequate?

C. Subtest reliability
   1. Are reported coefficients adequate?
   2. Are subtests sufficiently long?

VI. Norms (when appropriate)
   A. Are the norms provided adequate and appropriate?
   B. Does the normative sample seem truly representative?
      1. Up-to-date?
      2. Age—does the sample cover age range in your class?
      3. Sex—were appropriate numbers of boys and girls included?
      4. Socioeconomic level—does sample cover SES found in your class?
      5. Range of ability—does the sample include ability levels found in your class?
   C. Are the norms relevant for your class and purposes?

VII. Bias
   A. Is there evidence in the test content of sex stereotyping?
   B. Is there evidence in the test content of racial or ethnic bias?

VIII. General evaluation
   A. What is your general evaluation of the test?
   B. For what purpose and group would you recommend this test?

What Types of Tests are Available?

The final consideration with respect to test selection concerns information about the types of tests available for the teacher. (See Chapter Five for a more detailed discussion of types of tests.) Most reading tests can be classified under four headings: reading achievement and/or survey tests, informal tests, diagnostic tests, and readiness tests. These types differ with respect to what they attempt to measure.

Standardized Norm Referenced Reading Achievement and/or Survey Tests

Reading achievement or reading survey tests are defined as tests which provide an approximate level at which a child is functioning in relation to a specified group. Such tests are usually

standardized in terms of administration procedures and normative information. These tests usually measure word recognition, vocabulary, comprehension, and, in some cases, rate of reading. Some may measure more competencies but all usually attempt a general achievement score which reflects overall reading performance.

The diversity of test titles among available reading achievement tests raises two questions which may influence test selection. First, although test titles may differ, do the tests agree on the tasks they require students to perform? Second, do tests which purport to measure similar abilities measure them in the same way?

An examination of standardized achievement and survey tests of reading at the primary level reveals a fair degree of uniformity in the reading dimensions measured; usually word recognition and comprehension are most often measured. However, differences emerge in the labels used to identify the dimensions and the manner in which they are measured.

Tests which measure word recognition, at the primary level, may also be called vocabulary, word knowledge, or word meaning on some tests. In addition, there are a number of different ways word recognition can be measured. For some tests, the task is selecting the word associated with a picture from among four choices. Another type of task requires the reading of brief incomplete sentences prior to selecting a word which completes that sentence. Some other tests require students to select from several printed choices the word which has been pronounced by the teacher. This task is quite different from the matching or selection tasks of the first two examples since both auditory and visual discrimination are required for successful completion. Yet, all three tasks may carry similar labels.

The content of reading comprehension tests, at the primary level, also reflects different types of items. Some tests require students to examine several pictures and mark them in accordance with directions given in a sentence. Another variation is to present a single picture and require the selection of an appropriate sentence from among several which accompany the

picture. A third type requires the child to read a short paragraph and demonstrate understanding by recognizing what was read in translated form. Again, each of these tests will probably have a similar title but all use different techniques to assess comprehension. A question arises as to whether the tests are really measuring the same facets of comprehension.

A similar situation exists with respect to standardized achievement and survey tests used in the intermediate grades. While word knowledge and comprehension are again the basic components, there is disagreement concerning the labeling of the subtests; and the measurement tasks are quite different from test to test, even though apparently the same ability is being measured.

Some differences do exist between primary and intermediate grade standardized reading achievement and survey tests. First, intermediate grade tests rarely use pictures; thus, students cannot rely upon picture clues. Second, subtests attempt to measure more specific areas of comprehension (finding main idea, determining cause and effect, predicting outcomes). Third, there is an increased emphasis on requiring inferences which go beyond the literal meanings stated.

Values and uses of standardized reading tests. Basically, these tests can provide data to help teachers answer the following five questions.

1. What is the general reading performance of a student or a group? A student or a group can be compared to the norm sample (using standard scores) or students in a group can be rank ordered, high to low.

2. What are the general areas of strengths and weaknesses for a student or a group? Since most of these tests give at least two subtest scores as well as a total score, subtest scores can occasionally indicate real differences in achievement between major skill areas if certain cautions discussed in Chapter Five ("percentile bands" and "standard error of measurement") are followed. Instruction can then be adjusted to reflect these discrepancies.

3. Which students in the group have the same general

needs? Standard scores can help teachers form groups within a class or help make placements to classes.

4. Is appropriate instructional material available? Students achieving above and below average probably need special instructional material or need the use of the materials adapted to their performance levels. Material may need to be bought, located, created, or adapted.

5. Has a student or a group made measurable progress in general reading performance? Comparing test results on two different occasions can help determine how much progress has been made during instruction. Nearly all good standardized tests provide at least two different forms for such use. [There are some problems in such a procedure. Interested readers should consult a measurement text such as *Measurement and Evaluation in Psychology and Education* (*23*).]

Limitation of standardized reading tests. The standardized norm referenced reading achievement and/or survey tests are integral parts of most reading programs. They possess a number of positive values and uses previously described for the reading program. However, it should be realized that these tests have certain limitations as well.

1. The very fact that a test is "standardized" in terms of administration and scoring may make it inappropriate for use with certain individuals or groups. The test may be too difficult or too easy; items may be meaningless or placed at inappropriate levels; directions may be difficult to understand.

2. The test maker's quest for brevity may result in unrealistic time limits and a choice between depth and breadth in sampling. The scores of children who work very slowly but accurately may be suspect; the sampling of behavior may be superficial or constricted.

3. Group administration may work to the disadvantage of certain individuals. The group situation combined with the standardizing conditions may invalidate the test results in some instances.

4. The format of the test may restrict the type of items used. A machine scorable format, for example, virtually

demands some form of multiple-choice items. Certain abilities may not be adequately sampled with multiple-choice items.

5. The quality of the information obtained about an individual student may be less specific than that obtained for groups of students. Therefore, caution should be exercised in making use of the scores when planning for individuals.

Criterion Referenced Testing

There has been much debate on the degree to which norm referenced tests can help teachers make instructional decisions. It has been suggested that norm referenced tests result in little information about what a student can do. For example, what does a raw score of 121 points, a grade score of 4.2, or a percentile of 63 mean with regard to the actual reading tasks a student can do? Scores like these usually are useful for decisions of a comparative nature rather than revealing specifically what a student can and cannot do.

What are criterion referenced tests? If present standardized reading tests do not meet instructional needs, what approaches exist for developing assessment procedures to meet these needs? Criterion referenced measurement has been suggested as a useful alternative. A student's score is interpreted by comparing it to some absolute standard, usually some specified behavioral criterion of performance; the focus is on what the student can do. For example, a criterion referenced phonic analysis test may state that a score above 80 percent is necessary for adequate learning to be assumed. What other students score is not important. The question is whether the student has demonstrated a required level of proficiency in a skill and, thus, has reached a specified goal of instruction.

Criterion referenced tests are tied to the concept of mastery: the purpose of such a test is to measure achievement of a specific behavior in an either/or situation. For example, has June mastered the essential readiness skills to move on to formal instruction? Has Lila mastered the skills necessary for critical reading? Can Phillip read orally with 80 percent word recognition accuracy? In each situation, the criterion is quite

definite and the individual student is assessed to determine whether the task can be completed successfully. In other words, the prespecified objectives and levels of achievement either have or have not been attained.

Criterion referenced tests often resemble commercial diagnostic tests and their results are often used for nearly identical purposes. Most diagnostic reading tests are usually administered by a special reading teacher only after other reading tests have been given or after teacher observation raises the possibility of some rather severe reading problems, while many criterion referenced tests are used routinely by classroom teachers as an ongoing part of the total instructional program.

Limitations of criterion referenced tests. Several interrelated assumptions appear to be necessary in the criterion referenced approach. Although there is some debate about these assumptions, they include:

1. Mastery (a student either has or has not achieved the objective(s) satisfactorily) is a reasonable criterion.
2. Each item in a test has inherent worth by design.
3. A hierarchy or sequence of skills and knowledge exists in any skill or content area.
4. There is common consensus regarding the skills selected for testing.

Some limitations of criterion referenced measurement include:

1. Objectives involving hard-to-measure qualities, such as appreciation, attitudes, or inferential reading comprehension, may be included on the test and the resulting scores may be suspect.
2. Objectives involving the retention and transfer of what is learned may become secondary to the one-time demonstration of mastery of stated objectives.
3. Specifying the domain of tasks (determining critical instructional objectives) is of extreme importance. Good tests will do nothing to overcome the problem of poorly developed objectives.
4. Determining proficiency standards can be troublesome.

Perfect or near-perfect performance should be required if a) the criterion objective calls for mastery, b) the skill is important for future learning, and c) items are objective type and guessing is likely.

5. Test items frequently represent skills in too isolated a context rather than in an applied and/or functional situation.

Which type of measurement is best? Teachers need not choose between norm referenced and criterion referenced measures. The two types of measures ought to complement each other, with each type chosen according to the purpose for testing. If instruction is individualized so that time is a variable and a student keeps at a task until mastery results, then criterion referenced testing or teacher made tests are appropriate. On the other hand, if instruction is structured so time is constant, then students will be expected to achieve at different levels, and norm referenced measures should be used to discern differential achievement.

The teacher and criterion referenced tests. Reflection on criterion referenced measurement may make teachers wonder whether the idea is new. The answer is "Not really." Teachers have always had instructional objectives and have assessed students to determine whether those objectives have been achieved. The increased emphasis on criterion referenced measurement is in the availability of commercial tests for formal assessment in much the same way that norm referenced tests are now used. Also, many people are encouraging teachers to write criterion referenced tests for use in their own classrooms. This aspect is discussed in Chapter Three.

The teacher of reading should be most concerned with the concept of criterion referenced testing as it is used in the variety of informal tests which are used as part of the reading program. It is in assessing daily or weekly progress that the notion of criterion referenced measurement has the most utility. Here the teacher is concerned with how well a student is doing in terms of the stated instructional objectives. In essence, this is criterion referenced measurement.

Informal Reading Measures

In planning and assessing a reading program, teachers will often find it necessary to seek information that is not available from existing tests or to supplement information from them. When this is the case, it is often necessary to devise informal measuring instruments. Informal measurement can be the heart of diagnostic teaching since it can be done regularly and frequently in the classroom setting and can be used to quickly and effectively emulate a number of classroom reading situations.

What are various types of informal reading measures? Following are examples of some of the most useful informal techniques for assessing reading ability, particularly regarding strengths and weaknesses in specific skill areas.

Teacher observation. This is the most naturalistic informal technique. It is often overlooked but is a viable means by which teachers can diagnose student reading skills in the classroom. An alert, skillful teacher has countless opportunities during the day to gather much information about student progress and needs by observation and, thus, gain real insights into the problems a child may be encountering when faced with a particular reading task.

Observation can be most useful for evaluation of word recognition, comprehension, oral reading fluency, and interest in and attitude toward reading. Through observation, answers to the following questions about a student's progress in learning to recognize and use words can be determined:

1. Does the student retain words previously introduced?
2. Has there been an increase in ability to use phonic and/or structural elements in words?
3. Is there improvement in the ability to understand word meanings?
4. Has the ability to analyze and pronounce words independently increased?
5. Has there been an increase in oral reading ability?
6. Is spelling ability improving?

With respect to comprehension, information can help answer the following questions:

1. Can materials be read with greater ease, speed, and comprehension than previously?

2. Has there been improvement in the ability to follow directions for completing class and independent assignments?

3. Has skill in retelling materials read increased?

4. Has the ability to recall specific facts from a reading selection improved?

5. Has there been improvement in the ability to answer thought-type questions based upon reading materials?

Interest in and attitude toward reading can also be assessed through observation. It should be directed by the following questions:

1. Does the student read independently more eagerly than in the past?

2. Has the quantity of independent reading increased?

3. Has the student developed special reading interests?

4. Does the student read a wider variety or higher quality of materials than previously?

Some teachers find observation difficult because they do not know what to look for. The checklist in Chapter Four may serve as a practical means for systematizing observations.

Teacher observations can be further systematized by use of anecdotal records. The simplest form consists of a folder where dated observations, gathered in either a systematic or incidental manner, are kept in chronological order. In addition, samples of a student's work collected along with the anecdotes can be valuable. The primary purpose is to aid the teacher in assessing the developing reading characteristics of the child. Without readily available checkpoints, gradual but steady improvement may be mistaken for no improvement. They may also be superb for communication with parents, other teachers, administrators, and support personnel.

Informal tests. Many basal reading series include non-standardized tests which can be used for quick checks of students' word analysis skills, comprehension abilities, oral reading behavior, or any associated skill. (Teacher made criterion referenced tests, discussed in Chapter Three, are examples of

informal tests.) Such tests can also be constructed by the teacher to assess students' understanding of just-presented material or to obtain samples of various kinds of reading behavior.

Bliesmer (*3*) has suggested several "do's and don'ts" for teachers in order that they may make more effective use of informal testing in teaching reading:

DO	*DON'T*
1. do try to test informally in settings where children can share application of skills in realistic and functional situations	1. don't make every testing situation a formal occasion
2. do make sure materials used for informal testing are of appropriate levels of difficulty	2. don't assume materials are too easy just because a child answers all or most questions over the material quickly and correctly
3. do try to find out why children responded as they did	3. don't assume that a child who has marked or written a "wrong answer" has necessarily answered incorrectly
4. do minimize the writing required to answer or respond to test items	4. don't try to make testing an attempt to check on all the various reading skills
5. do use questions which call for exercise and exhibition of various thinking skills and which get at various levels of comprehension	5. don't test mainly for the recall of facts
6. do be sure that materials permit a child to show or practice the skills which are supposedly being checked	6. don't assume that almost any material will be appropriate for checking almost any skill
7. do remember that a number of and the testing of them are basically oral matters	7. don't try to do all testing through silent reading

Informal measures are useful in providing continuous assessment of student performance. Since it more closely parallels the actual reading situations the child encounters in the classroom, an informal test can provide the information necessary for prescribing immediate corrective instruction.

Informal reading inventories. Many teachers use an informal reading inventory (IRI) to measure a student's oral and silent reading. Since standardized tests of reading may not be

accurate for placing children in instructional materials, informal reading inventories offer a solution to the problem of which book best fits a child. IRIs can be used to determine a student's functional reading levels.

The three functional reading levels include:

1. Independent level—the level at which the student can read fluently without teacher assistance. The student can read material with no help and virtually perfect performance. The usual criteria for the independent level is 99 percent accuracy in word recognition and 90 percent or better comprehension.

2. Instructional level—the level at which the student can make maximum progress under teacher guidance. It is the level where the student is challenged but not frustrated. To be considered at the instructional level, materials should be read with 95 percent word recognition accuracy and, at least, 75 percent or better accuracy in comprehension.

3. Frustration level—the level at which the student can no longer function effectively. The student is unable to deal with the material and manifests a variety of behavioral characteristics to indicate the difficulty, including refusal to read, lack of expression in oral reading, excessive lip movements during silent reading, word pronunciation difficulties, word-by-word reading, and others. In addition to the behavioral characteristics, the formal criteria are word recognition of below 90 percent and 50 percent or less comprehension.

Johnson and Kress (*12*) provide detailed instructions for constructing an informal reading inventory. In addition, there are prepared graded selections for use in constructing informal reading inventories. However, in addition to informal reading inventories which can be constructed by the teacher, there are "formal" informal reading inventories. They are almost identical to teacher made IRIs and contain graded word lists and graded paragraph selections with questions to test comprehension. The major distinguishing characteristic lies in the fact that formal IRIs are published commercially. Neither type of IRI is free of difficulties. Problems concerning the reliability, the adequacy of the comprehension questions, and others are found in teacher made IRIs, but the content validity of the instrument is usually

satisfactory for forming instructional reading groups. Commercially available IRIs usually do not have relevant technical data available, although it can probably be assumed that the instrument was carefully constructed.

Diagnostic Reading Tests

While standardized achievement or survey tests can be used to locate strengths and weaknesses, questions about specific aspects of reading performance are best answered by diagnostic type tests. If students seem to have unusual difficulty with a particular phase of reading or if there seems to be a wide discrepancy in the scores made on different sections of a reading achievement test, then a diagnostic test should be administered. They are usually given only after other tests have been given or teacher observation raises the possibility of some rather severe problems. These tests usually yield more subtest scores than an achievement or survey test. A diagnostic test or battery may include measures of oral reading, comprehension, and word analysis skills including sight vocabulary, visual and auditory discrimination, rate of reading, or reading-study skills. Special teachers rather than classroom teachers will probably select, administer, score, and interpret these tests because of their specialized nature but the classroom teacher may see the results or be asked to adjust instruction in light of the interpretation of these results.

Data from diagnostic tests should answer the following questions:

1. In what specific skills may a student be deficient?
2. Is the student having difficulty with understanding the basic skills, or is the problem mainly that of skills application?
3. Is there a discrepancy between the student's knowledge of phonic rules and facility in using them?
4. In light of the testing, which of the student's weaknesses should receive attention initially?

What do reading diagnostic tests really diagnose? In attempting to answer the question, "What do diagnostic reading tests really diagnose?" Winkley (*25*) studied nine diagnostic

reading tests. It was concluded that no common definition of a "diagnostic reading test" was found. Several other problems were identified which suggests that caution should be exercised in using and interpreting diagnostic reading tests. However, these problems should not preclude their use since they do present a more systematic analysis of reading skills requiring assessment than procedures devised by a classroom teacher. The reading diagnostic test can make a more comprehensive analysis of reading abilities which can then be supplemented with referral measures specifically designed for individual students.

Choosing a diagnostic reading test. How does a teacher choose a diagnostic test? The same principles which are utilized in selecting other types of tests are appropriate for diagnostic tests. The factors of validity, reliability, and usability are again most important.

The validity of each subtest should be determined to see whether they represent meaningful areas for providing remedial instruction. If subtests are too global, then they do not provide much more information than what is available from a survey test. Also, while the same label may be used on several different subtests, they may measure entirely different abilities.

Reliability should be primarily assessed with regard to whether it is sufficiently high for individual use. Because a variety of subtests is found in diagnostic tests, the teacher must determine whether the subtest is long enough to be sufficiently reliable since the longer the subtest the more likely it is to be reliable. A high level of reliability is required when interpreting and comparing subtest difference scores of individuals. Teachers should not assume that just any difference between subtest scores is diagnostically meaningful. Standard errors of measurement (see Chapter Five) require that there be considerable differences between subtest scores before teachers conclude that there is a real difference in achievement between the subtest categories. An example is given under "percentile bands" in Chapter Five.

In addition to the above factors, Ramsey (*20*) has suggested other desirable criteria for diagnostic tests:

1. The reality criterion—Does the test assess an ability in much the same manner as the ability is used in real reading?

2. The guessing criterion—Does the test make it impossible for the student to guess the correct answer to an item?

3. The specificity criterion—Does an item measure a specific ability rather than a constellation of abilities?

4. The active criterion—Does the response demand some overt, observable behavior so that the examiner can clearly discern the nature of the student's difficulty?

5. The comprehension criterion—Do items that assess understanding of what is read actually test comprehension and interpretation rather than pure memory of what has been read?

Thus, for those pupils who are experiencing difficulty in reading, an instrument to pinpoint the nature of the problems contributing to poor reading performance is required so remedial instruction can be appropriately planned. The diagnostic reading test is designed to fill this need.

Reading Readiness Tests

Another type of test which many teachers will encounter is the reading readiness test. In a majority of elementary schools, reading readiness tests are administered and considered important components of the beginning reading program.

What is reading readiness? Reading readiness remains a controversial topic because of the issue of whether readiness is an intrinsic state of the student or whether it is the accumulated experience of the individual. Traditionally, individuals attempting to define reading readiness have viewed it as the result of maturation. Recent knowledge, however, has suggested that a student's readiness to learn to read is a product of both maturation and environmental factors. Within this framework, reading readiness is defined as the capacity to profit from systematic reading instruction.

This view of readiness, however, is incomplete since an integral component of the readiness process is a consideration of the types of learning opportunities which will be made available. This suggests that readiness is related to a set of circumstances in the form of methods, materials, and the level at which instruction begins.

What do reading readiness tests measure? The typical reading readiness test is intended to help select students who are ready to begin formal, systematic reading instruction.

Predictive validity of readiness tests. The teacher should seek evidence that the test accurately predicts which students will learn to read best. Do students who score high on a readiness test become good readers? And vice versa? The evidence generally suggests that, on the whole, readiness scores alone have only a fair degree of precision in predicting reading achievement for individual students.

Does this mean that reading readiness tests are not valid predictors? MacGinitie (*13*) called attention to two considerations which suggest that reading readiness tests can be useful. First, the predictive validity of present readiness tests is comparable to other instruments used for the purpose of prognosis in achievement. Second, readiness tests do appear to measure factors which have a relationship with first grade reading achievement. This is especially true if the tasks used to assess readiness closely resemble the criterion, that is, reading behavior. The reasons for the predictive validity of tasks that closely resemble reading are 1) they are more likely to require similar kinds of learnings, 2) they measure the past influence of environmental and motivational factors that are likely to continue to influence in a similar way the further acquisition of reading skills, and 3) what a student already knows cannot fail to be learned. Thus, the more closely factors in readiness tests resemble actual reading behavior, the higher the relationship between the readiness test and subsequent reading achievement.

Teachers should be aware that test scores obtained from preschool children are frequently highly unreliable. This means that test scores are subject to rapid change based on the effectiveness of school learning situations and developmental changes. Therefore, judgments must be tempered accordingly. For example, it is often helpful to have loose rather than rigid grouping of students for reading in primary grades based on kindergarten test results.

Factors in choosing a reading readiness test. The following are some considerations for using readiness tests:

1. Reading programs should be analyzed to determine the skills and abilities most important for success in beginning reading. For example, a teacher utilizing a whole word or look-say approach, would probably consider visual discrimination of letters and words a critical skill. Conversely, a strong phonics emphasis in beginning reading would suggest auditory discrimination as a crucial factor for success.

2. It should be realized that not all the important readiness factors are measurable with standardized tests. Therefore, teachers can enhance predictions by combining readiness tests with observation and evaluation of factors such as oral language facility, experiential background, attitude toward reading, and rate of learning in classroom situations.

Conclusion

The purpose of this chapter has been to provide guidelines and information relevant for test selection. This is vitally important for a successful reading program because the selection of tests is one of the more neglected aspects in the planning of a testing program. Test selection, at first glance, appears to be a routine procedure that any teacher can handle with little preparation. On the contrary, the wise selection of tests to be used is an important phase of a testing program which calls for considerable understanding if reading tests are to be servants of the reading program.

References

1. Barrett, T.C. "Predicting Reading Achievement through Readiness Tests," in J.A. Figurel (Ed.), *Reading and Inquiry*. Newark, Delaware: International Reading Association, 1965, 26-28.
2. Berg, P.C. "Evaluating Reading Abilities," in W.H. MacGinitie (Ed.), *Assessment Problems in Reading*. Newark, Delaware: International Reading Association, 1973, 27-34.
3. Bliesmer, E.P. "Informal Teacher Testing in Reading," *Reading Teacher, 26* (1972), 268-272.
4. Buros, O.K. (Ed.). *Reading: Tests and Reviews*. Highland Park, New Jersey: Gryphon Press, 1968.
5. Buros, O.K. (Ed.). *The Eighth Mental Measurements Yearbook*. Highland Park, New Jersey: Gryphon Press, 1978.

6. Farr, R., and N. Anastasiow. *Tests of Reading Readiness and Achievement: A Review and Evaluation.* Newark, Delaware: International Reading Association, 1969.
7. Farr, R. *Reading: What Can Be Measured?* Newark, Delaware: International Reading Association, 1969.
8. Glock, M. "How the Classroom Teacher Can Use a Knowledge of Tests and Measurements," in R.E. Leibert (Ed.), *Diagnostic Viewpoints in Reading.* Newark, Delaware: International Reading Association, 1971, 31-40.
9. Harmer, W.R. "The Selection and Use of Survey Reading Achievement Tests," in T.C. Barrett (Ed.), *The Evaluation of Children's Reading Achievement.* Newark, Delaware: International Reading Association, 1967, 53-64.
10. Hayward, P. "Evaluating Diagnostic Reading Tests," *Reading Teacher, 21* (1968), 523-529.
11. Hieronymous, A.N. "Evaluation and Reading: Perspective '72," *Reading Teacher, 26* (1972), 264-267.
12. Johnson, M.S., and R.A. Kress. *Informal Reading Inventories.* Newark, Delaware: International Reading Association, 1965.
13. MacGinitie, W.H. "Evaluating Readiness for Learning to Read: A Critical Review and Evaluation of Research," *Reading Research Quarterly, 4* (1969), 396-410.
14. McDonald, A.S. "Using Standardized Tests to Determine Reading Proficiency," *Journal of Reading, 8* (1964), 58-61.
15. Merwin, J.C. "Educational Measurement of What Characteristic, of Whom (or What), by Whom, and Why," *Journal of Educational Measurement, 10* (1973), 1-6.
16. Millman, J. "Criterion Referenced Measurement: An Alternative," *Reading Teacher, 26* (1972), 278-281.
17. Mitchell, A.S. "Values and Limitations of Standardized Reading Tests," in J.A. Figurel (Ed.), *Forging Ahead in Reading.* Newark, Delaware: International Reading Association, 1967, 163-167.
18. Otto, W. "Evaluating Instruments for Assessing Needs and Growth in Reading," in W.H. MacGinitie (Ed.), *Assessment Problems in Reading.* Newark, Delaware: International Reading Association, 1973, 14-20.
19. Prescott, G.A. "Criterion Referenced Test Interpretation in Reading," *Reading Teacher, 24* (1971), 347-354.
20. Ramsey, W. "The Values and Limitations of Diagnostic Reading Tests for Evaluation in the Classroom," in T.C. Barrett (Ed.), *The Evaluation of Children's Reading Achievement.* Newark, Delaware: International Reading Association, 1967, 65-78.
21. Robinson, H.A., and E. Hanson. "Reliability of Measures of Reading Achievement," *Reading Teacher, 21* (1968), 307-313.
22. Sipay, E.R. "A Comparison of Standardized Reading Scores and Functional Reading Levels," *Reading Teacher, 17* (1964), 265-268.
23. Thorndike, R.L., and E. Hagen. *Measurement and Evaluation in Psychology and Education,* Fourth Edition. New York: John Wiley and Sons, 1976.

24. Traxler, A.E., and R.D. North. "The Selection and Use of Tests in a School Testing Program," *The Impact and Improvement of School Testing Programs: The Sixty-Second Yearbook of the National Society for the Study of Education*, Part II. Chicago: University of Chicago Press, 1963.
25. Winkley, C. "What Do Diagnostic Reading Tests Really Diagnose?" in R.E. Leibert (Ed.), *Diagnostic Viewpoints in Reading*. Newark, Delaware: International Reading Association, 1971, 64-80.
26. Womer, F.B. "What Is Criterion Referenced Measurement?" in W.E. Blanton, R. Farr, and J. Tuinman (Eds.), *Measuring Reading Performance*. Newark, Delaware: International Reading Association, 1974, 34-43.

Chapter Three

Construction and Analysis of Reading Tests

Michael L. Kamil
Purdue University

The array of tests in reading seems to force a choice between norm referenced and criterion referenced tests. As an aid in understanding the differences between such tests, this chapter will compare and contrast items from norm referenced tests with items from criterion referenced tests for a group of reading skills. Then some of the ways that published tests can be supplemented by teacher constructed items and tests will be examined. Finally some techniques for writing items and constructing tests will be reviewed.

This chapter will cover five skill areas:

1. Reading readiness
2. Decoding
3. Word analysis
4. Comprehension
5. Vocabulary assessment

Several important areas of reading instruction have, for brevity's sake, been omitted from this chapter: work-study skills, content field reading, rate and flexibility, literary appreciation, and interest in and attitude toward reading. Readers interested in evaluating one or more of these areas should consult a methods textbook on the teaching of reading.

In order to choose or develop a test, needs and objectives must be carefully determined. Tests must always be closely related to objectives, since objectives should determine teaching content. As specific classroom objectives are clearly established, subskills in the broader areas (like those listed above) will be determined. Such specific objectives could then be translated into formal tests which could, in turn, be compared to items on a test. The degree to which the content of the test items matches the content of the objectives will determine the value of the particular tests for that specific application. There are other considerations that determine the value of a test. Among them is the use to which the results are put (discussed in Chapter Two).

We will examine items on published tests that might be appropriate to measure each of these skill areas. This is not intended to be an exhaustive treatment, but rather a sketchy outline of the more complete procedure. It is assumed that the entire range of objectives in each of these areas is available elsewhere.

Reading Readiness Skills

The skills that are necessary for a child to begin formal reading instruction typically include some or all of these: auditory and visual discrimination, following directions, listening comprehension, letter name knowledge, vocabulary, blending sounds together to make words, and reading words.

Figure 3.1 is an example of an item taken from the Metropolitan Readiness Test, a norm referenced test. This item, from the School Language and Listening subtest, requires a child to mark a picture that corresponds to the sentence the examiner reads.

Figure 3.2, from the Cooper-McGuire Diagnostic Word Analysis Test, a criterion referenced test, illustrates another way in which reading readiness is assessed. This item requires the child to be able to discriminate initial sounds. The test's authors feel that a score of 80 percent or better correct on this subtest indicates that the child has adequate auditory discrimination to begin formal reading instruction.

Figure 3.1

A. Put your finger on the little black BALL. Mark the picture that shows this: The duck is beside the flower... The duck is beside the flower.

Reproduced from the Metropolitan Readiness Tests. Copyright ©1974 by Harcourt Brace Jovanovich, Inc. Reproduced by special permission.

Figure 3.2
Test R3: Auditory Discrimination

Objective R 3: Given four pictures whose names begin with three different sounds, the learner will be able to mark two pictures beginning with the same sound as two dictated words with 80 percent accuracy.

Percentages Recorded:

%			
27	33	40	47
53	60	67	73
80	87	93	100

Sample Test Item: The pupil is presented with pictures of four common objects. The teacher identifies the pictures (e.g., "The names of these pictures are *letter, book, fish, light*"), and the pupil is asked to mark two pictures that begin with the same sound as two additional words pronounced by the teacher (e.g., *leg* and *lamp*).

Rationale: The ability to discriminate initial sounds is an important readiness step that leads directly into the more complex skills of identifying sounds with appropriate letters, which is presented in Test P1. The 15 sounds that were chosen for the test are important only insofar as they measure a general ability to discriminate sounds; therefore, percentages (instead of the particular sounds) are noted on the Class Record Chart. An 80 percent criterion level is sufficient for the establishment of this readiness skill. Further instruction in auditory discrimination is given in meeting Objective P1, and complete (100 percent) mastery is expected at that time.

Cooper-McGuire Diagnostic Word Analysis Test, J. Louis Cooper and Marion L. McGuire. New London, Connecticut: Croft Educational Services, Inc., 1972.

Teachers can also develop tests of reading readiness if the contents of published tests do not seem appropriate. For example, the uppercase letters of the alphabet could be written on cards and children asked to name the letters. Or auditory discrimination could be tested by asking children to listen to word pairs and then say whether they were the same or different (ran-man, play-plan, than-then).

Decoding Skills

These skills, often broadly called phonics, are related to readers' abilities to translate letter (or spelling) patterns into sound.

The Stanford Diagnostic Reading Test, a norm referenced and criterion referenced test, has items such as those in Figure 3.3. These items tap the ability of students to match a letter with the sound it represents.

In one example in the Wisconsin Design for Reading Skill Improvement, a criterion referenced test, the teacher reads a list of words such as *man* and the student names the vowel letter in each word. Items in this subtest do not seem to tap a reading skill directly. But these items are examples of construct validity (see Chapter Five), they predict which children will most likely have the highest end-of-first-grade reading achievement.

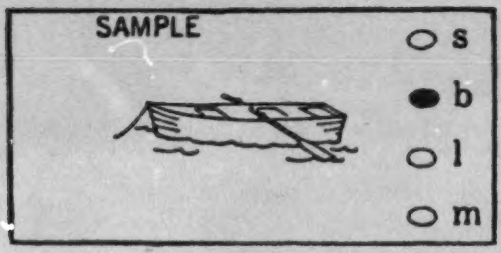
Teacher made items can be based on any letter-to-sound correspondence rule. Children could be given an example of CVCE patterns (as in the word *mate*) and asked to pronounce them. Some reading authorities suggest using nonsense words like *vate* to insure that children don't know the word by sight and must apply the rules. If a teacher is unable to find these items on a published test, this alternative would be useful.

Structural Analysis Skills

These skills are differentiated from decoding skills by the unit of analysis and include inflected endings, prefixes and suffixes, and syllabication skills.

The Stanford Diagnostic Reading Test, a norm referenced and criterion referenced test, uses items to test syllabication

Figure 3.3



Look at the shaded box at the top of the page where you see a picture of a boat. Think of the beginning sound of the word "boat"... "boat." What is the first sound in "boat"?

Pause for replies.

Yes, it is /b/ ("buh"). "Boat" begins with the /b/ sound. Which letter in the shaded box stands for the /b/ sound?

Pause for replies.

Yes, it is "b" ("bee"); that is why the space next to the "b" has been filled in in your booklet. You will do the same with the other questions on this page. I will tell you what the picture is. Then you will fill in the answer space in front of the letter or group of letters that stands for the beginning sound or sounds of the word I say. Mark only one letter or group of letters for each picture. Does everyone understand what to do?

Reproduced from the Stanford Diagnostic Reading Test, Red Level. Copyright ©1976 by Harcourt Brace Jovanovich, Inc., New York. All rights reserved

(Figure 3.4). In this type of item, the student is required to segment the *first* syllable of a printed word and match it with its visual representation.

The same test also requires the child to mark the syllable that is "extra" when two of the choices form a word. This item is from Part B, structural analysis, Stanford Diagnostic Reading Test and is designed to illustrate a student's ability to blend syllables into words (Figure 3.5). Blending of this sort is not precisely a structural analysis skill. The teacher must decide whether this subtest has the desired validity and must be certain that this information will be important for instructional decisions.

Figure 3.4

A tiptoe t ti tip

Look at line A in the shaded box at the top of the page. (Demonstrate.) You see the word "tiptoe." "Tiptoe" can be divided into parts (syllables) that can help you say the word. Look at the three groups of letters next to "tiptoe." Which group of letters is the *first* part (syllable) of "tiptoe?"

Reproduced from the Stanford Diagnostic Reading Test, Green Level, by permission. Copyright ©1976 by Harcourt Brace Jovanovich, Inc., New York. All rights reserved.

Figure 3.5

Part B

Samples			
A	my	so	self
B	mon	key	ber
1	air	one	plane
2	con	cor	test

Look at line A in the shaded box at the top of the page. Here you see three word parts. Two of these parts can be put together to make a real word. The other part is extra. The parts that make the word are in the correct order, from left to right. What word can you make from two of the three parts in line A?

Pause for replies. Say:

Yes, the word is "myself." What is the extra part? (Pause.) Yes, the extra part is "so." That is why the space under "so" has been filled in in your booklet. Now look at line B. Find the word that can be made from two word parts. (Pause.) What is the extra part?

Pause for replies. Then say:

Yes, the extra part is "ber." So fill in the space under "ber." "Mon" and "key" make "monkey," so "ber" is left over. Are there any questions about what you are to do?

Reproduced from the Stanford Diagnostic Reading Test, Green Level, by permission. Copyright ©1976 by Harcourt Brace Jovanovich, Inc., New York. All rights reserved.

The Cooper-McGuire Diagnostic Word Analysis Test measures the use of prefixes and suffixes as shown in Figure 3.6.

Figure 3.6
Test S6: Use of Prefixes and Suffixes

Objective S6: Given a list of prefixes and suffixes, the learner will be able to identify the affix to be added to a given root word to make sense in a sentence with 100 percent accuracy.

Sample Test Item: The pupil is presented with a sentence in which one of the words needs the addition of a prefix or a suffix (e.g., "When everything fell out of his suitcase, he had to ___pack it"). The pupil chooses an appropriate affix (one that satisfies the meaning of the sentence) from a list of prefixes and suffixes.

Prefixes

Meaning Units	Visual Units
anti-(against; the opposite of)	a-
circum-(around)	ab-
dis-(opposite of; apart)	ante-
ex-(out of; former)	com-, con-, col-
extra-(outside; beyond)	contra-
im-(not)	de-
in-(in; into)	inter-
in-(not)	per-
intra-, intro-(inside)	peri-
mal-(bad)	post-
mis-(wrong)	super-
non-(not; the reverse)	trans-
out-(more than; beyond)	be-
over-(too much)	en-
pre-(before)	
pro-(in front of)	
re-(again; back)	
self-(see dictionary)	
sub-(under)	
under-(below; not enough)	
un-(not; opposite of)	

Cooper-McGuire Diagnostic Word Analysis Test, J. Louis Cooper and Marion L. McGuire. New London, Connecticut: Croft Educational Services, Inc., 1972.

A specific instructional objective is given for each subtest of this test. This makes comparison with objectives of a particular instructional program quick and easy.

Teacher made items can be used to measure skills that are not found on commercial tests. Thus, a teacher who wants to know about children's knowledge of final syllables could create items analogous to those in Figure 3.4.

Vocabulary Skills

This broad category includes the meanings of words as well as the recognition of individual words.

There are two major types of tests for vocabulary skills, recognition and production. In recognition tests, a stimulus word is to be matched with its synonym from among several responses. Production tests require the student to make a response, not merely select an option.

Meaning vocabulary is measured normatively by recognition procedures on tests like the Gates-McKillop Reading Diagnotic Test (Figure 3.7).

Figure 3.7

Directions. Say to the child: "This is a test to find out how wel you know the meanings of words. I am going to read part of a sentence: 'A head is part of a _____.' Now I'm going to read four choices for finishing the sentence: 'coat, saw, man, box. A head is part of a coat, saw, man, box.' Which of the choices tells you what 'head' means?"

1. A head is part of a coat saw man box

Gates-McKillop Reading Diagnostic Tests, Forms 1 & 2, Arthur I. Gates and Anne S. McKillop. New York: Teachers College Press, Teachers College, Columbia University, 1962.

The skill involved here is the ability to recognize synonyms or attributes from a list of alternatives.

A production type task would be to ask a child, "In this selection, what does *responded* mean?"

Sight vocabulary tests emphasize recognition and pronunciation instead of word meaning. The Slosson Oral Reading Test gives measures of sight vocabulary in a norm referenced manner (Figure 3.8). This is a production task where the student must read the words in this list.

Figure 3.8

Directions:

1. Allow the child to read from one sheet while you keep score on another. At the start, say the following: "I want to see how many of these words you can read. Please begin here and read each word aloud as carefully as you can." (Indicate at what list to start.) "When you come to a

difficult word, do the best you can and if you can't read it, say 'blank' and go on to the next one."

List P

1. see
2. look
3. mother
4. little
5. here

Slosson Oral Reading Test (SORT), Richard L. Slosson. East Aurora, New York: Slosson Educational Publications, Inc., 1963.

The Johnson Basic Sight Vocabulary Test, a criterion referenced test, measures sight vocabulary (Figure 3.9).

Figure 3.9

Directions for Grade 1

In the row with a */* in front of it, find the word RED and draw a circle around it, *RED*.

1. reed red ready rod

From Form B of the *Johnson Basic Sight Vocabulary Test* by Dale D. Johnson. © Copyright, 1976, by Ginn and Company (Xerox Corporation). Used with permission.

The two examples of sight vocabulary measures differ in their objectives. One requires the student to produce a spoken response, the other requires the student to match a printed word with a word spoken by the teacher.

If teachers decide that the word samples used on published tests do not adequately match those of the instructional material, they may produce their own items. The results cannot be used indiscriminately, but this will be discussed thoroughly later.

Comprehension Skills

These are skills used by the reader to extract meaning from chunks of text larger than single words. Comprehension is most often measured by a series of questions following a passage. Figure 3.10 is an item taken from the Davis Reading Test, a norm referenced test.

Figure 3.10

Most people consider red salmon better than pink salmon, and it usually sells for more. A canner of pink salmon

printed on his labels, "Guaranteed not to turn red in the can."

8. What word is understood after "more"?
   A People
   B Uses
   C Money
   D Stores
   E Customers

Reproduced from the Davis Reading Test by permission. Copyright ©1960, 1961 by The Psychological Corporation, New York. All rights reserved.

An example of a criterion referenced comprehension test is the typical informal reading inventory (IRI) which accompanies many basal readers. If the child can meet the traditional criteria of 95 percent word recognition accuracy and 75 percent comprehension accuracy on a passage drawn from instructional material, it is assumed that the child can adequately handle that level material in a directed reading situation.

If teachers are unable to find sufficient diversity of question types in published tests to suit their needs, they can create their own questions. This is not as simple as it might first seem. Helpful guidelines can be found in *Classroom Questions* by Sanders and in *Teaching Reading Comprehension* by Pearson and Johnson. Additional suggestions are given later in this chapter in the section titled "Constructing Test Items."

More global measures of comprehension may be obtained by using cloze procedures for specific passages. The cloze procedure is a relatively simple test in which a student is required to fill in blanks in a passage of text. The results can be used normatively to rank a group of students or in a criterion referenced manner much as in an IRI in matching a child and instructional material.

The following steps can be used as a guide in producing cloze tests.

1. Select a passage of about 50-200 words, beginning with and ending with a complete sentence. The lower the readability level of the passage, the shorter it should be; for young students, passages should be kept short; for older students, they can be longer.

2. Leave the first sentence and last sentence intact and, beginning with the second sentence, retype the passage omitting every fifth word. Make the blanks all the same length.

3. Reproduce the passage and have the student fill in the blanks.

Once the student has filled in the blanks, score the passage by counting the number of *exactly correct* words filled in. Synonyms can be counted, but they won't change the rankings of students much and frequently it is very difficult to determine whether a word is acceptable.

Reading levels can be determined as follows: If a student scores 50 percent or better, the material is at the *independent level*; if he scores between 30-50 percent, the material is at the *instructional level*; below 30 percent, the material is too difficult and may be considered to be at the *frustration level*. (These levels are to be used as guidelines only. A student may score low for a variety of reasons. Also the cloze score is not readily generalizable to other materials.)

The following is a sample of a cloze passage of about second grade difficulty (*6*):

The Twin Rabbits

On the way to Ann's house, they saw two children going to the party. One was dressed as _____ white rabbit. The other _____ wearing a sailor suit.

_____ boy dressed as a _____ was standing at the _____ of Ann's house. He _____ the door for the _____ .

"That is Ann's _____ Tommy," Penny whispered to _____ .

Ann looked hard at the twin rabbits.

Because of the time, effort, and knowledge required to produce valid and reliable reading tests, it is strongly recommended that teachers make efforts to use commercially available tests. If a test meets your needs, use it rather than prepare your own. However, if you feel that published tests do not meet your requirements, prepare one. It would be unprofessional to try to make do with an inappropriate test.

Initial Considerations

Before beginning to construct a test, the following questions should be answered:

1. *What function is the particular test going to serve?* Is is going to be an integral part of a diagnostic teaching program; is it going to be a program evaluation; or will it serve some other function?
2. *Just what are the objectives that the test is designed to measure?* These may be specific instructional objectives or they may be general programmatic objectives.
3. *What types of test exercises will be most suitable?* The general choice is between objective and essay types of questions. Within the objective category, items may be true/false, multiple choice, matching, etc. Essay questions can require short answers or more involved, lengthier answers.

These decision processes are not easy. They require, first, that the teacher be intimately familiar with the educational goals and instructional objectives to be used. Second, the task of determining whether one's needs are met in a test takes time *and* effort. Finally, there is no guarantee a test can be found that will meet teachers' needs. In that case, teachers should use teacher made tests, which can be based directly on their own objectives.

The following are the things that need to be considered in analyzing or beginning to construct a test:

1. The types of instructional decisions you want to make.
2. The specific teaching philosophy and goals.
3. The amount of resources (time, money, personnel) to be used in evaluating or creating test instruments.
4. The amount of time to be spent in testing and/or evaluation.
5. The specific items on a test.

Constructing Objectives and Test Items

The first steps in writing test items are shown in the flow chart (Figure 1.1) in Chapter One. The chart shows that a needs assessment is necessary before specific instructional objectives can be formulated. *After* instruction, a student's performance is measured and evaluated. This section is concerned with the way in which tests and test items can be written by a teacher.

Writing Objectives

After a needs assessment has been conducted, general goals should be apparent. That is, some rather broad areas in which students need instruction should be evident. This section will first consider how to write specific objectives that can be used in instruction.

Generally, an objective should have two components: a specific behavior the student should be able to demonstrate and a level of acceptable performance. Thus, an objective like "The student will learn to read" is unacceptable. Neither the behavior nor the level of performance is specified. For a child learning to read, the objective would be improved if it were restated, "By the end of the school year, the child will read orally at the 1.9 grade level with fewer than 5 percent errors." Some readers may believe that this objective is still too general to be useful.

Here are some characteristics of objectives that will make them more useful:

1. *The objective should state a behavior that is observable.* Thus, an objective like "The students will be interested in reading" is poor because it is not really possible to observe "being interested in reading." An improved objective might be, "The students will voluntarily select and read five books of their own choice in x weeks' time."
2. *The objective should be specific and unambiguous.* The words used in the objective should not mean different things to different people. "The student will comprehend reading material at the fifth grade level" is not specific enough because comprehension has many meanings. A better choice might be, "The student will be able to correctly *paraphrase* passages at the fifth grade level of difficulty."
3. *Objectives should be unitary.* It is undesirable to cover more than one behavior in an objective. If two or more behaviors are in the same objective, it is difficult to specify the criteria. For example, this objective is poor, "The students will be able to *decode* and *define* words at

their reading levels." A better procedure is to have two objectives: 1) "The students will be able to decode words at their reading levels with 95 percent accuracy." 2) "The students will be able to define words at their reading levels with 80 percent accuracy."

4. *Objectives should be attainable under normal circumstances.* Given the amount of time allotted for instruction and the amount of resources a teacher can devote to instruction, the objective should be realistically attainable. For students reading two years below their grade placement, the objective, "The students will read at grade level at the end of this year," is probably unattainable.

Objectives should be constructed by first determining the desired student behavior, making sure that the characteristics mentioned above are not violated, even at the level of behavior. Performance should then be specified, preferably using active verbs like *to answer, to define, to say,* or *to locate.* Verbs such as *to feel, to think,* or *to grow* should be avoided. Finally, an outcome should be stated. The outcome should be an observable or measurable performance and should include a statement of the task and a level of mastery or a criterion to be achieved. For example: "The students will be able to categorize known words with the long or short sound of *oo* into two groups with 90 percent accuracy." The behavior is to classify words with *oo* in them. The students must distinguish between the long and short sounds of the letter combination. On a test of this skill, students must obtain a score of 90 percent before attaining that objective.

Notice that each objective specifies, in itself, a test. If a student satisfies the objective, instruction should move to new objectives. If a student does not meet the criterion, revisions in instruction, in the objective itself, or in the needs assessment may be necessary. A complete discussion of this process appears with the flow chart presented in Chapter One.

Writing objectives is probably much easier for certain enabling reading skills like decoding. More global skills like comprehension are fairly difficult to adapt to this purpose. That is, it is simpler to write test items that measure the ability to

pronounce the sound of a letter than it is to measure the understanding of even a short passage.

An important point to be derived from the discussion and examples above is that norm referenced and criterion referenced types of tests *can* be different, but they do not *have* to be different. When the comparison is between good instruments, the two types of tests are not very different.

Writing Items

There are some precautions that must be taken when writing tests and test items. Objective type items will be considered first, even though some of these warnings will apply to essay items and tests as well.

1. The most important single principle is that *the reading level and vocabulary difficulty should be within the range of the students who are going to take the test.* An item which reads, "separate the following list of words into two categories representing the same initial phoneme" is clearly beyond the reading abilities of most primary school students. It would be better to write, "Make two groups of words from the following list so that all of the words in each group start with the same sound."
2. *The items should be clear and unambiguous.* If you ask whether the words *bring* and *thing* sound alike, the question is ambiguous. It is better to ask whether the two words rhyme or end with the same sound.
3. *Test items should cover important content.* The more trivial the answer to a test item, the less effective it is in tapping a reading related skill.
4. *Items should be independent.* The answer to a question should not depend on whether the student has been able to answer a previous question correctly. If you are testing comprehension of a story, ask: "True or False, the events in this story occurred before 1863," rather than "True or False, the Civil War occurred before 1863."
5. *Every test item should have a single correct answer.* If

experts in the field can agree on the answer, it is appropriate. If there is some disagreement, the item should be discarded or revised. It is important to remember that an item should be constructed so that a student can get it correct *only* if the corresponding learning objective has been attained.

Test constructors favor the selection of items, for statistical reasons, that are missed by about 50 percent of the norming population. (Although, this is not the *same* 50 percent for each item). Items are also selected to maximize the overall reliability of the test. In addition, the items should be those that best predict the behavior that is being tested. Thus, it is not always readily apparent that an item is related to a particular behavior. Test constructors often are primarily concerned with predicting the presence (or absence) of a skill or behavior.

An example may serve to illustrate this last point. Discrimination is important in reading readiness. Readiness items often deal with colors, pictures, or abstract symbols. For norm referenced tests, these items might have been shown to be valid by correlating performance on them with performance on other, independent measures of reading readiness.

Criterion referenced tests often demand that items be more directly related to actual reading behaviors. That is, the items should involve words or letters rather than other pictures or symbols.

For many criterion referenced tests, a sequence of skill assessment *and* instruction is specified. In general, this is not the case for norm referenced tests. When test items relate directly to instruction, the sequence of assessment should also match directly.

For essay test items, some of the same criteria apply. The test must deal with important content. The task proposed must be specific. Above all, there must be a correct answer.

In general, it is easier to construct essay or short answer items; however, what is gained in construction is often lost in the additional complexity of scoring. Essay items are limited in the amount of content that can be sampled, whereas objective

questions can be used to cover more of the content. The units of learning you want to emphasize will also determine the type of test you choose. Essay tests cause students to focus on much larger, integrated chunks of content than do objective tests.

The Finishing Touches

When the objective items have been completed, some care must be taken with the format of the test before it is administered. Above all, it must be easy to read. Items should not be squeezed so close together that they are hard to distinguish. It is a good idea to group items of the same type together. That is, true/false items should not be mixed with multiple choice or matching items. When testing several different types of content in the same test, group together those items which relate to similar content. A test should probably start with the easier items and progress to the more difficult ones. This will allow the test taker to gain some confidence and help sustain motivation. Directions must be clear and specific. The reading level of directions must certainly be within the capabilities of the students taking the test. It is also important to avoid patterns of answers, e.g. alternating true and false answers (T, F, T, F, etc.) or sequencing multiple choice answers (ABCDCBA, etc.).

Before using a teacher made test, construct it and put it away for some time. Then, go back, read it over, and take it yourself. Have a colleague take the test, also. Revise or eliminate any items that fail to meet the criteria discussed, that you have difficulty with, or that seem odd. This may mean that more items will have to be written. Teachers shouldn't be discouraged if their efforts at constructing tests don't yield the results wanted. The procedures followed by commercial test producers are very involved, very expensive, and highly specialized. It is not suggested that teachers can replace all tests with those that they make themselves, but rather that they can supplement commercial tests with teacher made tests for their own specific goals.

References

1. Cooper, J. Louis, and Marion L. McGuire. *Cooper-McGuire Diagnostic Word-Analysis Test.* New London, Connecticut: Croft Educational Services, Test Manual, 1972, 4-5, 18-19.

2. Davis, Frederick B., and Charlotte Croon Davis. *Davis Reading Test*, Form 2A, Series 2. New York: The Psychological Corporation, 1961, 2-3.
3. Gates, Arthur I., and Anne S. McKillop. *Gates-McKillop Reading Diagnostic Tests*, Forms 1 & 2. *Pupil Record Booklet*, 15; *Manual of Direction*, 10-11. New York: Teachers College Press, Teachers College, Columbia University, 1962.
4. Johnson, Dale D. *Johnson Basic Sight Vocabulary Test*, Forms A & B. *Directions Manual*, 8-9; student copy, 4-5. Lexington, Massachusetts: Ginn, 1976.
5. Karlsen, Bjorn, Richard Madden, and Eric F. Gardner. *Stanford Diagnostic Reading Test*, red and green levels. *Manual for Administering and Interpreting*, 20-21 (green level); 25-27 (red level). *Student Booklet*, 7-8 (red level), 7-8 (green level). New York: Harcourt Brace Jovanovich, 1976.
6. McKee, Paul M., et al. *Come Along*, level 2[1]. Hopewell, New Jersey: Houghton Mifflin, 1966, 205.
7. Nurss, Joanne R., and Mary E. McGauvran. *Metropolitan Readiness Tests*, Level I, Form P. *Teacher's Manual: Directions for Administering*, 26; *Metropolitan Readiness Tests*, 17. New York: Harcourt Brace Jovanovich, 1976.
8. Otto, Wayne. *Wisconsin Design for Reading Skill Development*. Minneapolis, Minnesota: National Computer Systems, 1972, 31.
9. Pearson, P. David, and Dale D. Johnson. *Teaching Reading Comprehension*. New York: Holt, 1978.
10. Sanders, Norris, *Classroom Questions*. New York: Harper and Row, 1966.
11. Slosson, Richard L. *Slosson Oral Reading Test* (SORT). East Aurora, New York: Slosson Educational Publications, 1963.

Chapter Four

Organizing and Applying Test Results

George Canney
University of Idaho at Moscow

There is an old saying, "Someone who jumps from a plane without a parachute is likely to jump to a conclusion." Teachers who do not exercise judgment and knowledge of the students they teach run the risk of "jumping to conclusions" that are no more satisfactory or instructionally sound than are those of the ill-fated airplane passenger. Teachers should interpret test data in light of daily observations of students' reading, while always remaining ready to question discrepant pieces of information. If teachers accept test scores as *truth*, whether or not the scores fit with their perceptions of individual students, then they are unlikely to be successful with students who really need help. Test scores are to be viewed as one source of information on student progress; to be meaningful for instructional purposes, teachers will have to examine scores carefully and keep any decisions tentative—always tentative.

How to Use Test Scores

Measuring student reading ability always results in some inaccurate information. However, by following the procedures suggested below, poor decisions based upon incomplete or inaccurate test results can be avoided. These guides are represented by the acronym VERIFI:

*V*alidate the content of the test items,
Keep *E*quilibrium in the assessment program by using different types of testing procedures.
*R*eview student progress frequently in order to check judgments about instructional procedures.
*I*nvestigate discrepancies between test scores and inclass performance.
*F*ormulate an instructional plan with students and share the general results of tests with them.
*I*nterpret all test data in light of what is known about the student and curriculum objectives.

Willingness to VERIFI information from tests, so that student progress is clear both to you and to students, will affect instructional success. Without VERIFIcation, test scores will be of little value and may even decrease the effectiveness of instruction. A more detailed review of these six steps may explain how they relate to effective teaching.

Validate. Examine test items carefully. Be certain that the test items actually reflect the content and skills you want to measure.

Otherwise, scores are meaningless for instructional purposes, and further item examination subsequent to testing is useless because you do not plan to present that content or skill.

You will want to study the items missed by a student. A low score may be the result of a single type of misconception, rather than a general confusion in the areas assessed.

Preparing a chart similar to the one presented in Figure 4.1 can assist in identifying test items that proved easy or difficult in order to make decisions about which skills or materials to present next. In this example, students were given a teacher-constructed test of their ability to identify five prefixes (+ refers to correct responses, - refers to incorrect responses).

From the patterns of correct and incorrect responses on this prefix assessment, it appears that Shawn is probably confused about all of these prefixes while Willie, Tom and Debbie can recognize in- and re- but have difficulty with sub-,

Figure 4.1
An Item-response Chart Showing Correct (+) and Incorrect (-) Responses on a Test of Prefixes.

Objective	Recognizing common prefixes											
Prefixes	in		re		sub		dis			un		
Items	1	4	6	12	2	9	3	8	10	5	7	11
Willie T.	+	+	+	+	-	-	-	+	-	+	+	-
Susan S.	+	+	+	+	+	+	+	+	-	+	+	-
Daphene J.	+	+	+	+	+	+	+	+	+	+	+	+
Tom B.	+	+	+	+	+	-	-	-	-	-	-	-
Shawn P.	+	-	-	-	-	+	-	-	-	-	+	-
Debbie M.	+	+	+	+	-	+	+	-	-	+	+	-
Ryan M.	+	+	+	+	+	+	+	+	+	+	+	+

dis-, and un-. Susan, Daphene, and Ryan have performed somewhat better than the other members of the group. But the entire group seems unsure about the prefixes dis- and un-. Another possibility is that items 10 and 11 may not be satisfactory items, since only two students answered them correctly.

Although time-consuming, procedures like the one just described are necessary if group administered tests are to be useful for making instructional decisions. This is especially true when using commercial tests which may contain items on content not yet presented, or tests which require students to respond differently from the way they ordinarily answer test questions.

Equilibrium. Keep a balance in your assessment program. Each form of measurement, especially observation, can contribute information useful for instruction.

If you use a standardized achievement test, consider the following procedures to assist in creating instructional groups.

Arrange the class test booklets in rank order, using reading comprehension scores. Form three groups of test scores approximating the top quarter, middle half, and bottom quarter of the class. On a separate sheet of paper, record the students'

names and reading comprehension scores. Then examine these scores. First, what is the range of scores and does this range represent a typical distribution for the school; is it broader or narrower, or higher or lower than usual? Such information will allow you to compare this class with other groups from your teaching experiences. Next, compare the three groups of test scores. Are they fairly close to each other, or distinctly separate? If they are close, the class may be fairly homogeneous in reading ability and you may want to consider grouping students quite flexibly. If the groups are quite different, and the students within the groups are also different from one another, you will definitely want to consider more than three reading groups.

An examination of other subtest scores, such as vocabulary, word attack skills, or rate, may help to estimate the homogeneity of the reading groups you have proposed. If discrepancies exist between a student's subtest scores (reading comprehension and reading rate or vocabulary), you may want to concentrate your instruction in those areas that appear weak. Students who achieve similar scores on a particular subtest may be grouped temporarily in a special skills group, even though their general reading performance differs widely.

To determine whether score discrepancies are significant, create bands or ranges of scores by adding the standard error of measurement, plus and minus, to the student's percentile rank score. If the score bands created for each subtest overlap a standard error of measurement by half or more, it is unlikely that the student's performance is really stronger in one skill than in the other. Less overlap suggests that the student *may* be stronger in one skill and requires special instruction in the areas deemed weak. Similar comparisons can be made for a group's subtest scores, created by averaging across the students' scores in the group. In this way you can obtain some estimate of group instructional needs in addition to the needs of individual students.

At this point the decision on how much further testing is needed before instruction can begin depends on how low the students scored relative to the expected achievement levels of

your school. You will want to learn as much as possible about students having difficulty in reading before placing them in materials that may be too difficult. When starting, it is better to be conservative than to overestimate a student's reading level. Improper grouping may lead to student frustration, loss of motivation, and subsequent instructional or behavioral problems.

You also want to learn more about the reading levels and strengths of the other children in your class; but since you have limited time to conduct needed testing for the low scoring children, you can probably place the more able students in materials and initiate instruction. Even if you place a student in an inappropriate group, your observations and the student's daily performance should lead to rapid adjustments in instruction.

Depending on the number of students you believe need further testing, you may now decide to use either a group diagnostic reading test or an individual one. Administering the individual diagnostic test is usually more time consuming, especially if you are using it with several students; but it will normally provide a more detailed picture of reading performance than a group test. Both types of tests often allow some latitude in using parts of the test battery, but the individual diagnostic test allows you to terminate some subtests if they prove too easy or too difficult for a particular student. The administration of specific subtests may be spread over several days. Your observations of the impact of careful instruction during this period, together with the student's performance on the subtests you administer, should help you adjust instruction to the student's abilities and interests.

If you wish to assess student abilities to read material from the specific texts used in your classroom, you may also employ an informal reading inventory (IRI) or your own test to gather highly specific information on each student's reading skills.

By coupling the use of test instruments with informal observations of student performance in both testing and learning situations, you bring an equilibrium, or balance, to the assessment program in reading.

Review. Review student progress frequently in order to monitor instructional effectiveness.

Once you have formed *tentative* judgments about each student's strengths, weaknesses, and interests, plan to observe individual progress frequently. Otherwise, it is difficult to check the accuracy of initial estimates of each student's reading level or monitor growth over time as a result of teaching and student effort. Quizzes, tests at the end of units, daily performance on assignments, conferences, and regular checks of oral reading performance (ordinarily used for beginning readers) will help you in making estimates of student needs.

Frequently, the most helpful method for monitoring student growth is diagnostic teaching. This is when students are assessed on the same material covered on a systematic daily basis. Oral reading for diagnosis is one assessment procedure that meets this criterion. On a frequent basis, once or twice a week, each student should read aloud from a new story in the reading text. As the student reads, record miscues (significant errors) on a separate copy of the story. Usually two minutes of reading (from 75 words for beginning readers to 200-word passages for better readers) provides an adequate sample to assess oral reading performance. Unless a student is particularly self-conscious about reading aloud in a group, oral reading for diagnosis can be done as a regular part of a group reading activity. Teacher proficiency in using oral reading for diagnosis comes with practice.

By using the marking system shown in Figure 4.2 and tape recording the student's oral reading, a useful record of student strengths and weaknesses in oral reading can be obtained. The tape recording permits rechecking; it also can be kept and replayed later to help students recognize their own progress in reading.

Pikulski (1974) notes in an article on informal reading inventories (IRI) that writers in the field of reading disagree on what oral responses constitute errors or miscues. Some writers do not consider omissions, substitutions, repetitions, or insertions as errors unless they affect comprehension. Others prefer to

Figure 4.2
Code System for Oral Reading for Diagnosis

Example:

The little girl was standing near the stop sign. While she was waiting // for the bus, she watched the squirrels playing in the trees. A big bird hopped from branch to branch. The little girl forgot to watch for the bus. While she was looking at the animals in the tree, the bus went past her. So the little girl began to cry.

Error Type	*Code*	*Example*
Omission	◯	little
Insertion	^	to near ^ the
Reversal	∿	spot stop
Substitution	\	next near
Repetition	←	girl was standing
Hesitation	//	waiting // for
Teacher Assisted	T	squirrels
Self-corrected	✓	saw watched
Disregards Punctuation	⌒	sign. While

examine patterns of errors to ascertain the strategies that a student is using to read, rather than to make a simple tally of those errors.

A second way of frequently reviewing progress is by using multiple response or every pupil response (EPR) techniques (*1*, *2*). Those procedures permit you to assess each student's performance on word attack skills and vocabulary. They are

designed for use in a group setting, since the students use the EPR devices to respond nonverbally to questions.

There are numerous variations of the EPR idea; two will be described here. Figure 4.3 depicts two EPR devices. In Example One a tongue depressor has been glued to a circle of tagboard 6 inches in diameter. Holes cut large enough for a child to insert a finger are arranged around the perimeter of the tagboard and numbered. The numbers correspond to the lines on a pocket chart (Example Two). The pocket chart, when made small enough (9"x12"), can be an individual student's EPR device. Students point to particular word cards on their pocket charts in response to the teacher's questions.

As an example of how the EPR technique works, suppose that nine previously introduced words, listed below, have been put onto phrase cards and arranged on the teacher's pocket chart as shown in Figure 4.3.

line 1	would	line 6	mountains
line 2	brake	line 7	shake
line 3	pleased	line 8	goats
line 4	cold	line 9	flint
line 5	replace	line 10	_____

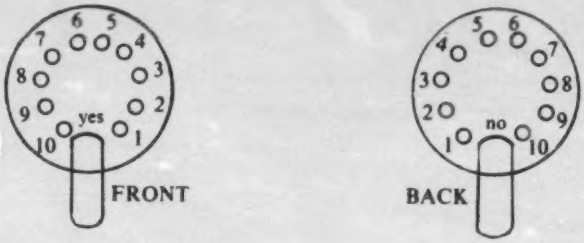
After reading through the words once, the students are given a handheld EPR device and instructed to answer each of the questions you ask. They are to respond by inserting one or more fingers through the numbered holes in the EPR device that correspond to the numbered lines on the teacher's pocket chart. If the students are using their EPR pocket charts, each student has a miniature pack of nine words arranged on his pocket chart in the same order as found on the teacher's pocket chart. The students will simply point to the appropriate words on their pocket charts.

With the students arranged facing you so that you can see each one clearly, begin asking prepared questions:

1. Does the word on line seven mean "an old cabin"? (word meaning)
2. Which word has the same sound in the middle as the

Figure 4.3
Every Pupil Response Devices for Maximizing Practice in Reading

Example One: Hand-held EPR device



Example Two: EPR Pocket Chart

would	mountain
1	6
brake	shake
2	7
pleased	goats
3	8
cold	flint
4	9
replace	
5	10

word /sit/?
(auditory analysis) 9

3. Which word means the opposite of hot?
(word meaning-antonyms) 4
4. Which word has an ending that means more than one?
(structural analysis-inflected endings) 6,8

More than one correct response to a question is sometimes possible. This encourages students to be somewhat

more independent in making decisions about the correct answer. It also provides an opportunity for students to explain why their answers may differ.

The EPR procedures are useful for reviewing student progress because they provide a brief, fast-paced means to observe how much each student in a group knows about the various skills and vocabulary being presented. A student who continually selects the wrong answer or who frequently looks at peer choices is often the student who is less sure about the skills than other students in the group. Poor performance may also indicate that the instructional material is too difficult.

Of course, continued observation of student performance on daily assignments such as worksheets, independent projects, and choices of books during free reading periods contributes significantly to the review of each student's progress. However, unless a means is developed to record your observations of student performance, you may not remember each student's strengths and weaknesses in planning future lessons. The teacher's guide could become your record of what students know and have yet to learn.

Commercially developed recording systems are available, usually as part of skills management systems like The Wisconsin Design for Reading Skill Development (1973) or the Science Research Associates Mastery Kit (1974). Many basal reading programs also have management systems available.

One school developed their own progress chart (Figure 4.4). The chart reflects most word attack skills and could be expanded to include other areas of concern in reading. Skills could be added or deleted to better serve the instructional concerns of particular schools and teachers. Printed on a manilla folder sized cardboard, the chart can be used to hold recent assessment information useful for planning weekly assignments and creating special skills groups.

A modified checklist procedure is used to record student strengths. Where appropriate (letter/sound correspondence, etc.), the term *mastery* is used. Elsewhere, the terms *reads* and *spells* are employed. The teacher records the month and year

when a particular skill is *introduced* (I); then, when the student can read/spell that pattern well, the teacher checks *reads* (R)/*spells* (S). The chart accompanies the student from year to year, or from classroom teacher to reading specialist, to facilitate communication. Skills checked M (mastery), R (reads), or S (spells) are reviewed in later lessons.

Such charts allow teachers to review the progress of each student on specific skills. A blank cell might indicate the need for additional instruction, a special worksheet, or the creation of a special skills group to consolidate the student's knowledge in this area.

While it does take time to review student progress, there is no other way to match effectively your own teaching expertise and the materials you have at hand to the needs and interests of students.

Investigate. Examine discrepancies between test scores and daily class performance.

As part of the VERIFIcation process, compare student test scores with their inclass performance. Usually, the manuals of standardized tests provide estimates of the degree to which particular scores are unreliable or can change on a random basis from one testing experience to the next. This fluctuation of test scores is termed the standard error of measurement (SEmeas) and is normally provided for each subtest in the examiner's manual. Single scores should always be thought of as a midpoint within a range of possible scores. When interpreting scores, add or subtract 1 SE_{meas} to the single score point to obtain a range or score band. This will provide a more accurate picture of the student's true score (see Chapter Five for a more detailed discussion of Standard Error of Measurement).

Many factors combine to affect how a student performs on a test at any given moment. If you observe that a student's scores are lower than inclass performance would suggest, consider that nonacademic factors, such as fatigue or disinterest, may have contributed toward the lower test score. On the other hand, if the student's scores are higher than inclass performance

Figure 4.4

PUPIL PROGRESS CHART

GRADE: ____ TEACHER: _____
GRADE: ____ TEACHER: _____
GRADE: ____ TEACHER: _____
GRADE: ____ TEACHER: _____
GRADE: ____ TEACHER: _____
GRADE: ____ TEACHER: _____
GRADE: ____ TEACHER: _____

KEY

R - Reads D - Defines M - Mastery
S - Spells I - Introduced a - adequate
W - Writes

READING READINESS

	adequate
Directionality { Left-to-Right	
Directionality { Top-to-Bottom	
Visual Discrimination	
Visual Memory	
Auditory Discrimination	
Auditory Memory	
Kinesthetic Readiness	

SIGHT VOCABULARY

GRADE	adequate
1	
2	
3	
4	
5	
6	

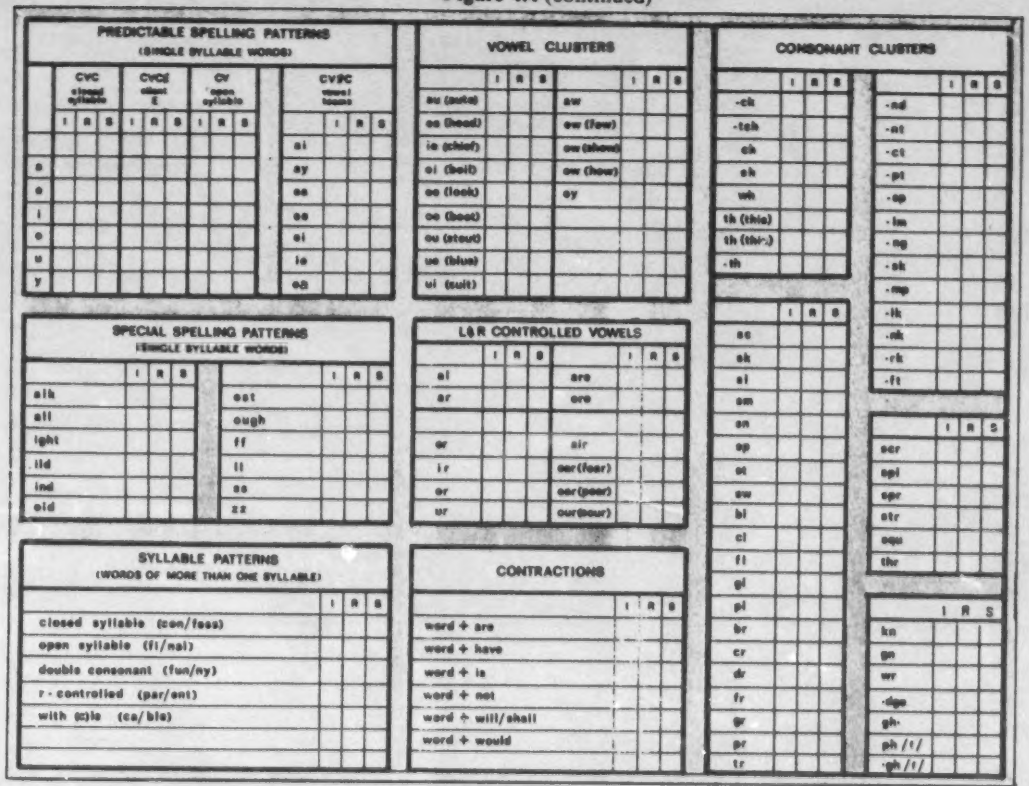
ALPHABET

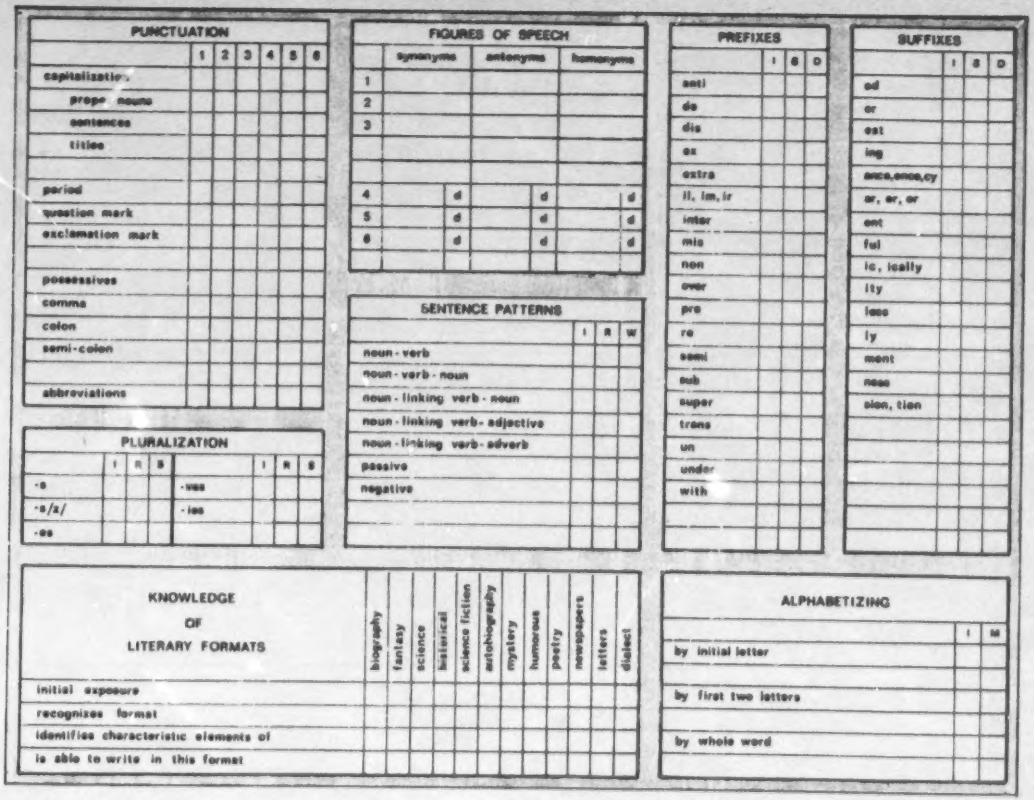
LETTER	Naming I	Naming M	Visual Discrimination I	Visual Discrimination M	Matching Sound/Letter I	Matching Sound/Letter M	Sound Production I	Sound Production M	Manuscript I	Manuscript M	Cursive Lower I	Cursive Lower M	Cursive Upper I	Cursive Upper M	
A															A
B															B
C															C
C (cat)															•
D															D
F															F
G															G
G (huge)															•
H															H
J															J
K															K
L															L
M															M
N															N
P															P
R															R
S															S
T															T
V															V
Q (U)															Q
W															W
X															X
Y															Y
Z															Z
E															E
I															I
O															O
U															U

STUDY SKILLS							
Reference Materials	K	1	2	3	4	5	6
1. Pictionary							
2. Dictionary							
locate words							
selects appropriate definition							
pronunciation keys							
guide words							
3. Encyclopedia							
locates subject							
guide words							
cross-references							
4. Card Catalog							
subject							
title							
author							
5. Thesaurus							
6. Atlas							
7. Filmstrips							
8. References by topic							
9. Glossary							
10. Bibliography							
11. Footnotes							
12. Maps							
13. Graphs							
14. Charts							
15. Takes notes from writings							
films							
lectures							
Reporting Skills							
1. Summarizes main idea of sentence							
paragraph							
2. Places events in proper sequence							
3. Identifies topic sentence							
4. Report Formats:							
oral							
composition							
book report							
summary							

SPECIAL TEACHER NOTES		
	major series used	comments
K		
1		
2		
3		
4		
5		
6		

Figure 4.4 (continued)





PUNCTUATION

	1	2	3	4	5	6
capitalization						
proper nouns						
sentences						
titles						
period						
question mark						
exclamation mark						
possessives						
comma						
colon						
semi-colon						
abbreviations						

FIGURES OF SPEECH

	synonyms		antonyms		homonyms	
1						
2						
3						
4		d		d		d
5		d		d		d
6		d		d		d

PREFIXES

	I	S	D
anti			
de			
dis			
ex			
extra			
il, im, ir			
inter			
mis			
non			
over			
pre			
re			
semi			
sub			
super			
trans			
un			
under			
with			

SUFFIXES

	I	S	D
ed			
er			
est			
ing			
ance, ence, cy			
ar, or, er			
ent			
ful			
ic, ically			
ity			
less			
ly			
ment			
ness			
sion, tion			

SENTENCE PATTERNS

	I	R	W
noun - verb			
noun - verb - noun			
noun - linking verb - noun			
noun - linking verb - adjective			
noun - linking verb - adverb			
passive			
negative			

PLURALIZATION

	I	R	S		I	R	S
-s				-ves			
-s/z/				-ies			
-es							

KNOWLEDGE OF LITERARY FORMATS

	biography	fantasy	science	historical	science fiction	autobiography	mystery	humorous	poetry	newspapers	letters	dialect
initial exposure												
recognizes format												
identifies characteristic elements of												
is able to write in this format												

ALPHABETIZING

	I	M
by initial letter		
by first two letters		
by whole word		

would have led you to expect, examine the types of activities you are using in your classroom. It is possible that the test is not accurately measuring what has been presented or that the student is inconsistent in the ability to use the skills (or that you may have underestimated the student's reading ability).

Such efforts to *investigate* discrepant pieces of information about student performance will contribute significantly toward your overall teaching effectiveness.

Formulate. Consider instructional plans in light of test scores but include the student in some of the planning.

When test results are considered as an indication of progress and are presented as indicating strengths rather than weaknesses (skills learned rather than items missed), then they can be used positively. Most people tend to persist in something when they experience early and frequent success in it and to lose interest when their efforts repeatedly lead to failure. Since poor readers have usually experienced more than their share of failure, especially on tests, using test results in a positive manner to encourage renewed effort may prove beneficial. In addition to a conference with a student, a visual chart that describes progress may prove helpful in maintaining student interest and suggesting new instructional plans.

Interpret. Use test results to assist in determining the effectiveness of instruction and interpret all test data in light of your knowledge about the students and the instructional objectives.

There is no substitute for sound judgment. Teachers should know the instructional capabilities of their students better than anyone else. This does not mean, however, that teachers can afford to ignore evidence that is inconsistent with their beliefs. For example, Charlie was a student who regularly received Cs and Ds in school. What the teachers thought of his ability was never clear. In taking the Scholastic Aptitude Test in eighth grade, prior to entering high school, Charlie scored 30 points higher on the test than the "best" student in the class. What factors contributed to Charlie's poor performance in class? One

dismaying possibility is that his teachers "knew" that Charlie was a C and D student and, consequently, he was seldom expected or encouraged to do better.

The point is that people may slip easily into interpreting information in terms of their expectations. Teachers must always subject their interpretations to independent (outside) verification and to question any discrepancies they observe. In this framework lies the key to the successful use of test scores: your willingness to VERIFI [illegible] you think you know about your students.

Using Assessment to Guide Instruction

The information obtained through measures of reading performance, including daily observations, can be used to suggest the need for programatic changes and to guide instruction.

The purpose of assessment continues to be the determination of student progress in learning to read and the acquisition of evidence to guide programatic changes in light of that purpose.

In a single building, where the conventional unit of analysis is the classroom or a grade level, scores on standardized achievement tests or criterion referenced tests in skills management programs may be used by the principal or reading teacher to evaluate the school reading program. If large discrepancies in student performance appear, questions about particular classroom reading programs may require that the principal and teacher(s) confer about reading instruction in that building.

The principal might also be interested in identifying strengths and weaknesses in the reading program, especially across grade levels. For example, the reading program adopted for the primary grades may be developing superior decoding skills in the students so that reading achievement scores are above average. However, the principal may observe a sharp drop in reading test scores in fourth grade, suggesting that the primary reading program may not be balanced sufficiently to produce students who read text with comprehension, not simply decode print. In this setting, combining reading scores by grade level may

reveal serious deficiencies (or strengths!) in the school's reading curriculum and prompt the establishment of a special curriculum committee to study the reading program.

If test data are used to assess program effectiveness, these questions may prove helpful:

1. How do students compare in reading growth at the same grade level, ability level, and cultural/economic group? In this type of analysis, national, state, and district norms are especially helpful because they provide an independent standard for comparison.
2. How well have the students attained specific program objectives? While district norms for standardized achievement tests might be used, group diagnostic reading tests, criterion referenced tests, and teacher made tests are also helpful.
3. How well are students progressing this year in light of past achievement in reading? Parents and teachers alike are interested in this comparison. A marked change in reading growth might be a powerful indicator of program strengths and weaknesses, although other nonprogramatic factors could produce similar changes.
4. How effective is this program with students of different ability levels? Perhaps a special reading program is needed for the more able students, the slowest readers, or the culturally different students.

While these questions focus on reading program effectiveness, in the final analysis each student's progress must be of primary concern. No program has ever been found to be 100 percent successful; consequently, to teach diagnostically requires that teachers attend to each student's needs. Questions like these can help individualize instruction:

1. Is there a need to provide further instruction in this area?
2. Is further practice all that is necessary to consolidate the student's knowledge in this area?
3. Is a different instructional approach needed for this student?

4. Is there a need for materials that more appropriately match this student's interests and ability?
5. Is there some indication that the student belongs with another group?

In seeking answers to these questions, an examination of achievement tests, diagnostic tests, quizzes, assignments and general classroom performance should contribute useful information for you to assess student growth in reading. Teachers prepared to VERIFI the observations they make can expect to discover new ways to assist their students toward becoming independent and critical readers.

References

1. Gambrell, Linda, and Robert Wilson. *Focussing on the Strengths of Children.* Belmont, California: Fearon Press, 1972.
2. Manning, John. "Using Every Pupil Response Techniques to Manage Skills Instruction," personal communication, 1973.
3. Otto, Wayne, and Eunice Askov. *The Wisconsin Design for Reading Skill Development.* Minneapolis: National Computer Systems, 1975.

Chapter Five

Definitions of Measurement Terms

Fred Pyrczak
California State University at Los Angeles

These definitions are arranged topically, with topics arranged to maximize understanding for readers who wish to read the chapter in its entirety. The following alphabetical guide to the definitions is designed for those who seek information on specific terms.

Term	*Page*
Construct	78
Content Validity	78
Criterion Referenced Test	73
Criterion Related Validity	78
Diagnostic Test	74
Domain Referenced Test	74
Face Validity	79
Grade Equivalent	76
Group Test	75
Individual Test	75
Norm Referenced Test	73
Percentile Band	76
Percentile Rank	75
Reliability	80
Standard Error of Measurement	80

Stanine 77
Survey Test 75
Validity 77

Types of Tests

Norm referenced test—a test that is tried out with a sample of examinees called a norm group so that an individual's performance can be described relative to that of the group. Often, the norm group consists of a carefully selected sample drawn from across the nation.

Norm referenced tests permit us to make statements such as this one: "Roger's performance was considerably better than that of the majority in the norm group, which consisted of students at his own grade level." Assuming that there were no special considerations in Roger's case, we probable would conclude that Roger's reading ability is above average. (See "Criterion Referenced Test.")

Criterion referenced test—a test designed to measure the extent to which an examinee can perform a desired behavior. A criterion referenced test usually is used to determine whether a student has reached a desired performance standard, hence the word *criterion*. The standard is determined independently of how a sample of examinees may have performed on the test. Usually, the standard is determined on the basis of the importance of the skill(s) being tested to the next skill in a hierarchy of skills or on the basis of the importance of the skill(s) to crucial, everyday activities.

A criterion referenced test usually measures a cluster of closely related skills. Often, the items in such a test are derived directly from a homogeneous group of instructional objectives. Criterion referenced tests usually have this characteristic since performance standards are easier to establish for tests confined to a limited number of objectives than they are for wide ranging tests.

The characteristics of a criterion referenced test may be understood by considering this example. Suppose we built a test

on initial consonant sounds. The content of the test is limited in scope and refers to a clearly identifiable and restricted reading objective. We could establish a performance standard in terms of numbers of items that must be marked correctly before a student is ready to move on to the next instructional objective, which might cover consonant blends. Unfortunately, there is no simple mechanical rule for establishing a performance standard for a test of this type. The test constructor or teacher would have to establish the standard in light of what is known about the importance of the skill tested to the next skill in the instructional sequence; in the absence of objective information on this matter, the standard would have to be established on a more or less subjective basis. Note that the performance of some group of examinees, such as other students in the class, would not be used to establish the standard.

Domain referenced test—a test in which the tasks are based upon a systematic analysis of specific subskills underlying a broad skill area called a domain. To the extent that we can identify all of the subskills and associated processes that underlie the domain of reading, we can build a domain referenced test by systematically sampling subskills to be covered in the test. Note that all good tests are based upon careful planning of skills to be measured. The crucial difference between a domain referenced test and other types of tests is the degree to which the tasks for a test are based upon a systematic sample of subskills and processes thought to underlie the entire domain.

Development of domain referenced tests is a relatively new endeavor. To reduce the scope of the task, some researchers are concentrating on the development of domain referenced tests that cover only one or more portions of the domain of reading.

Diagnostic test—a test that measures reading skills that are thought to be independent and yields a separate score for each skill. Diagnostic tests are based upon the notion that there are discrete reading skills; the portion of a diagnostic test that measures a given skill often is called a subtest. For example, the Diagnostic Reading Scales by Spache yields separate scores for

Word Recognition, Oral Reading, Silent Reading, and eight subtests in Phonics. Analysis of the pattern of scores indicates an examinee's areas of strength and weakness (see "Survey Test").

Survey test—a test that measures a variety of reading skills but yields only one score to represent an examinee's overall performance. In practice, few tests are pure survey tests since most yield two or more scores that may be compared in order to diagnose areas of strength and weakness (see "Diagnostic test"). Tests which measure only a limited number of broadly defined reading skills, however, often are referred to as survey tests. For example, the Davis Reading Test commonly is referred to as a survey test even though it yields separate scores for level-of-comprehension and speed-of-comprehension.

Group test—a test designed to be administered to a group of examinees at one time. Most group tests present examinees with written materials as well as oral directions; examinees are asked to respond by making marks on an answer sheet or in a test booklet (see "Individual test").

Individual test—a test designed to be administered to only one student at a time. Obviously, individual tests are less efficient than group tests because fewer students can be tested in a given amount of time. They are more flexible, however, for three reasons. First, they can be used to test skills (such as oral reading ability) that cannot be tested in a group situation. Second, when test directions permit, we can ask additional questions to be sure that a student understands the tasks being presented. Finally, we can adjust the beginning and ending level of a test according to a student's ability. For example, if the test directions permit and if the items are arranged in order of difficulty, we may stop testing once an examinee has missed a given number of consecutive items (see "Group test").

Types of Scores

Percentile rank—a type of score that indicates the percent of examinees in a norm group that scored below a given point.

Juan's percentile rank of 79, for example, indicates his performance on a test was better than that of 79 percent of the examinees in the norm group. The norm group for a test usually consists of a national sample of examinees at a given grade level. (See "Norm referenced test.")

Percentile band—a range of percentile ranks for an individual that helps us adjust our interpretation in light of random errors. We know that most scores deviate somewhat from what we would obtain if we could measure with complete precision. For example, if Bertha obtained a percentile rank of 89 on a vocabulary subtest and the band for her score ranged from 86 to 92, we would know with a specified degree of confidence that if we could measure with perfect precision, she would probably obtain a score somewhere between 86 and 92. The degree of confidence usually is 68 percent; that is, chances are 68 out of 100 that her true score is between 86 and 92.

When bands are given for two or more subtests, we can use them to interpret subtest score differences. For example, if Bertha also obtained a percentile rank of 81 on a comprehension subtest and had a band that ranged from 75 to 87, we would see that the bands for vocabulary and comprehension overlap. When bands overlap, we may conclude that the difference between the two performances may not be of practical importance in light of the lack of precision associated with the test.

Grade equivalent—a type of score that indicates the grade level (expressed in months of a grade) at which the average examinee in a norm group had a given number right. For example, Sarah's grade-equivalent of 5.7 indicates that the number of items she marked correctly was equal to that of the average norm-group examinee enrolled in the seventh month of the fifth grade.

Grade equivalents do *not* indicate the appropriate grade placement for an examinee. To understand why this is so, suppose that Sarah was enrolled in the third grade and took a test containing second, third, and fourth grade material. Her grade equivalent of 5.7 cannot possibly mean that she can do fifth grade

work since she was not tested on fifth grade material. It merely tells us that she can perform second, third, and fourth grade tasks as well as the average students in the seventh month of the fifth grade can perform these below grade level tasks. Even if Sarah's grade equivalent had fallen within the grade level range covered by the material in the test, it would not indicate appropriate grade placement for her since her score would represent some unknown combination of successes and failures on tasks for three different grade levels. Because of possible misinterpretations of grade equivalents, the most recent edition of *Standards for Educational and Psychological Tests and Manuals* calls upon test publishers to abandon or discourage the use of grade equivalents. For comparing a student's performance with that of a norm group, other types of scores such as percentile ranks are preferred.

Stanine—a type of score that places students into one of nine categories based upon the performance of a norm group. A stanine of "1," for example, indicates that the examinee's performance was roughly equivalent to that of the bottom four percent of those in the norm group. The following figure shows the meanings of stanines.

Stanine	1	2	3	4	5	6	7	8	9
Percent of Norm Group	Bottom 4%	7%	12%	17%	Middle 20%	17%	12%	7%	Top 4%

Stanines are useful when coarse groupings of examinees are desired.

Essential Characteristics of Tests

Validity—usefulness of a test for a given purpose. A reading readiness test designed to predict reading achievement, for example, is valid to the extent that it makes good predictions. A reading achievement test designed to measure second grade achievement is valid to the extent that is measures representative reading skills covered in instruction. Since the particular skills taught in two different second grade classrooms may vary, a test may be a valid measure of achievement in one classroom but not the other.

Note that a test may be valid for an illegal or unethical purpose. A test that discriminates against members of a minority group is a valid test for a test user who seeks to discriminate; it is invalid for one who wishes to measure and evaluate without discrimination. Hence, any discussion of validity must deal with the suitability of the purpose for which a test will be used.

Validity is sometimes divided into four major types:

Construct validity—the goodness of a test judged in terms of the extent to which the scores yielded by the test are meaningful in light of our theories and common sense notions. For example, a reading test that consistently failed to show gains for students exposed to systematic, intensive, and long term reading instruction (across different programs and different groups of examinees) would be said to lack construct validity, given that we have reason to believe that reading instruction does affect reading ability.

Content validity—The goodness of a test judged in terms of the appropriateness of the content (and skills) covered by the test.

Careful test planning is the key to insuring a high degree of content validity. Most commercial test publishers go to great pains to draw up outlines of content and skills commonly taught at specified grade levels and use them as a basis for writing test items. Teachers can produce tests with a high degree of content validity by drawing representative samples of objectives used for instruction and writing one or more test items per objective.

Criterion related validity—the goodness of a test judged in terms of the extent to which scores correlate with (put examinees in the same rank order as) some independent measurements of the trait being measured. For example, some reading tests have been validated by correlating the test scores with independent ratings of reading ability made by teachers.

When a test is administered at about the same time as the independent measurements are obtained, we say that we are estimating the test's *concurrent validity*, a type of criterion related validity. If we administered a test at about the same time teachers assigned grades in reading, for example, and correlated the test scores with teachers' grades, we would be estimating concurrent validity. When the test is administered at a time substantially in advance of the collection of independent measurements, we say that we are estimating the test's *predictive validity*, another type of criterion related validity. If we administered the test at the beginning of the first grade, and correlated the scores with teachers' grades assigned at the end of first grade, we would be estimating predictive validity.

The degree of criterion related validity possessed by a test usually is expressed with a correlation coefficient, a statistic with a range from -1.00 to +1.00. When used for expressing criterion related validity, the coefficient usually is known simply as a "validity coefficient." In practice, validity coefficients are positive in value. A coefficient close to "one" indicates high validity; a coefficient close to "zero" indicates low validity. Validity coefficients larger than .75 are rare.

Face validity—the goodness of a test as judged by its appearance. Usually, face validity refers to the goodness of a test from the point of view of people who are not trained in measurement. Face validity is not a scientific concept. It is important for tests to have face validity, however, in order to promote public acceptance of tests and testing.

Because there currently is a great interest in criterion referenced testing, individuals working on the theory for such tests have recently developed new, specialized notions of validity. Most can be classified in terms of the major categories given above. For example, for criterion referenced tests based upon specific instructional objectives, we may examine "congruence

validity"—the extent to which independent judges can match test items with objectives when they are given the items in random order. This type of validity may be thought of as a specialized type of "content validity."

Reliability—consistency of scores. The following illustrates one of the meanings of "consistency" when we are talking about norm referenced tests:

> The vocabulary knowledge of two students was measured with a test. James was found to have the most knowledge; Stan was found to have the least. A few days later, vocabulary knowledge was measured again with the same test. Again, James was found to have the most knowledge, and Stan was found to have the least. The two sets of results were consistent; the test was reliable. Note that we have not determined how useful (valid) our measurements are for a given purpose (see "Validity").

Reliability of norm referenced tests is expressed as a "reliability coefficient," a statistic with a range from -1.00 to +1.00. In practice, reliability coefficients are positive in value. A value close to 1.00 indicates high reliability; a value close to 0.00 indicates low reliability. Reliability coefficients of .80 and above are common and desirable for norm referenced tests. There is a variety of ways to estimate the reliability of norm referenced tests, and each has a different name. Hence, test manuals will use terms such as "split-half reliability," "parallel forms reliability," and "Kuder-Richardson reliability."

The reliability of criterion referenced tests is also a matter of concern, but the reliability of such tests is not usually expressed as a reliability *coefficient*. Other types of information should be provided—such as the percentage of examinees who may be misclassified as reaching the performance standard as a result of lack of precision in the test. Unfortunately, such information often is lacking for currently available criterion referenced tests.

Standard error of measurement—(abbreviated S_{meas} or SE_{meas}) is a statistic that tells us how many score points we should

"allow" for random errors that result from factors such as guessing and the emotional and physical status of examinees. For example, if Chuck obtained a score of 65 on a test with a standard error of measurement of seven points, he would probably obtain a score somewhere within seven points of 65 (58-72) if we could measure without error. Our allowance would be sufficient in about two-thirds of the cases; that is, we have about 68 percent confidence that the allowance is large enough. Therefore, in order not to deceive ourselves as to the accuracy of a score, it is best to think of a score as probably occurring within a band of scores rather than at a specific point.

The standard error of measurement is related to the reliability of a test. Other things being equal, the higher the reliability, the smaller the standard error of measurement will be. The relationship is not perfect, however, since the variance (differences) in test scores must also be taken into account in computing S_{meas}.